Evolutionary targeted discovery of influenza A virus replication inhibitors

Hershna Patel

Submitted to the University of Hertfordshire in partial fulfilment of the requirements of the degree of Doctor of Philosophy

August 2017

ABSTRACT

Influenza A is one of the most prevalent and significant viral infections worldwide, resulting in annual epidemics and occasional pandemics. Upon infection, antiviral drugs targeting the neuraminidase protein and M2 protein are the only treatment options available. However, the emergence of antiviral drug resistance is concerning, therefore the aim of this work was to identify inhibitor molecules that may bind to highly conserved regions of selected internal influenza A proteins. Sequences of the non-structural protein 1 (NS1), nuclear export protein (NEP) and polymerase basic protein 2 (PB2) from all hosts and subtypes were aligned and the degree of amino acid conservation was calculated based on Valdar's scoring method. Missing parts of the experimental structures were predicted using the I-TASSER server and ligand binding hot spots were identified with computational solvent mapping. Selected binding sites in conserved regions were subjected to virtual screening against two compound libraries using AutoDock Vina and AutoDock 4. Two out of twelve top hit compounds predicted to target the NS1 protein showed capability of reducing influenza A H1N1 replication in plaque reduction assays at concentrations below 100 µM, although the target protein and mechanism of action could not be confirmed. For the NEP, conservation analysis was based on 3000 sequences and binding hot spots were located in common areas amongst three structures. Docking results revealed predicted binding affinities of up to -8.95 kcal/mol, and conserved amino acid residues interacting with top compounds include Arg42, Asp43, Lys39, Ile80, Gln101, Leu105, and Val109. For the PB2 protein, conservation analysis was based on ~12,000 sequences and fifteen potential binding hot spots were identified. Docking results revealed predicted binding affinities of up to -10.3 kcal/mol, with top molecules interacting with the highly conserved residues Gln138, Gly222, Ile539, Asn540, Gly541, Tyr531 and Thr530. The findings from this research could provide starting points for *in vitro* experiments, as well as the development of antiviral drugs that function to inhibit influenza A replication without leading to resistance.

ACKNOWLEDGEMENTS

I thank Dr Andreas Kukol (principal supervisor) for help and guidance to complete this work. I also thank Dr Ralph Rapley (second supervisor) for support and advice, Jamie Stone and Prof Martin Hardcastle for technical assistance, Dr Angie Lackenby and Claudia Rosenow (Public Health England, Colindale) for allowing me to use their facilities and for help with antiviral experiments, and Noelia Perez-Diaz, Giulio Fusco, and Ashish Patidar for sharing their expertise with Western Blotting.

This work has made use of the University of Hertfordshire high performance computing facility and was funded by the School of Life & Medical Sciences. I am grateful to the department of Biological and Environmental Sciences for funding my studentship, and extend my thanks to all staff members that have supported me throughout my time here.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Å	Ångström
CPSF30	Cleavage and polyadenylation specificity factor 30
CRM1	Chromosome region maintenance protein 1
DARTS	Drug affinity responsive target stability
DMEM	Dulbecco’s modified eagle medium
DMSO	Dimethyl sulfoxide
ED	Effector domain
GTP	Guanosine triphosphate
HA	Haemagglutinin
IFN	Interferon
LDS	Lithium dodecyl sulfate
M1	Matrix protein 1
M2	Matrix protein 2
MD	Molecular dynamics
MDCK-SIAT	Madin Darby canine kidney 2,6-sialyltransferase
MEM	Minimum essential medium
NA	Neuraminidase
NCBI	National centre for biotechnology information
NCI	National cancer institute
NEP	Nuclear export protein
NLS	Nuclear localisation signal
NS1	Non-structural protein 1
PA	Polymerase acidic
PABP II	Poly A binding protein II
PAGE	Poly acrylamide gel electrophoresis
PAINS	Pan assay interference compounds
PB1	Polymerase basic protein 1
PB2	Polymerase basic protein 2
PBS	Phosphate buffer saline
PDB	Protein data bank
PME	Particle mesh ewald
RBD	RNA binding domain
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
RNP	Ribonucleoprotein
ROCE	Receiver operator characteristic enrichment
SDS	Sodium dodecyl sulphate
TNC	Tris Sodium Calcium
TPCK	Tolylsulfonyl phenylalanine chloromethyl ketone
VMD	Visual molecular dynamics
VTM	Virus transport medium

1 INTRODUCTION

1.1 The influenza A virus

There are three main types of influenza (flu) virus: A, B and C, which cause a contagious respiratory illness in humans characterised by fever, cough, sore throat, headache and muscle pain. Infection with type A and B virus strains cause seasonal flu, which peaks during the winter months and ranges in severity depending on the immunocompetence level of the host. Infection with influenza C is much less common and is associated with mild illness only (Taubenberger & Morens, 2008).

Besides human to human transmission via close contact or inhalation of virus droplets, influenza A displays zoonosis with wild aquatic birds being the natural reservoir (CDC, 2014). Therefore, the virus has the potential to be transmitted between various hosts such as birds, pigs, horses and whales amongst many other animals. Such cross-species transmission has lead to the formation of novel re-assortant influenza A subtypes through mixing of viral genome segments (antigenic shift), conferring a dangerous increase in pathogenicity to the human population (Taubenberger & Kash, 2010). In addition to annual epidemics which may result in significant death rates, serious outbreaks have been reported in previous years due to reassortment events. This includes the 1918 Spanish flu pandemic killing over 40 million people, the 1957 Asian flu pandemic, the 1968 Hong Kong flu pandemic and more recently, the 2009 Swine flu pandemic which resulted in approximately 284,000 deaths worldwide within the first year of virus circulation (Dawood et al., 2012; Taubenberger & Kash, 2010). Influenza types B and C have a limited host range and do not cause pandemics; hence this work focuses on the discovery of antivirals targeting influenza A.

Influenza A viruses are subtyped numerically according to antigenic properties of the hemagglutinin (HA) and neuraminidase (NA) surface proteins as H*x*N*y*, and to date, there are 18 different HA subtypes and 11 different NA subtypes identified (CDC, 2014). The standard influenza A nomenclature consists of the antigenic type (e.g., A, B, C), host of origin, (for human-origin viruses, no host is given), geographical origin, strain number, year of isolation, and the HA and NA antigen description in parentheses, e.g. A/duck/Alberta/35/76 (H1N1). All subtypes are able to infect birds, with the exception of H17N10 and H18N11, which have only been discovered in bats (Wu et al., 2014). The species affected by different HA and NA subtypes are shown in table 1. As certain subtypes are associated with poor clinical outcome in humans, influenza A infection is considered a major medical concern to public health.

Table 1. Species affected by the different HA and NA influenza A subtypes, with wild aquatic birds being the main reservoir (table adapted from CDC, 2014). Both HA and NA proteins of the subtype must match to a host as they recognise the same cell surface receptor in order to establish infection (Mitnaul et al., 2000).

HA SubType	Humans	Poultry	Swine	Bats/ Other	NA SubType	Humans	Poultry	Swine	Bats/ Other
H1	✓	✓	✓		N1	✓	✓	✓	
H2	✓	✓	✓		N2	✓	✓	✓	
H3	✓	✓	✓	other animals	N3		✓		
H4		✓	✓	other animals	N4		✓		
H5	✓	✓	✓		N5		✓		
H6	✓	✓			N6	✓	✓		
H7	✓	✓		other animals	N7	✓	✓		other animals
H8		✓			N8	✓	✓		other animals
H9	✓	✓	✓		N9	✓	✓		
H10	✓	✓			N10				Bats
H11		✓			N11				Bats
H12		✓							
H13		✓							
H14		✓							
H15		✓							
H16		✓							
H17				Bats					
H18				Bats					

Currently in the United Kingdom, a trivalent flu vaccine (usually composed of two type A strains and one type B) is formulated annually to protect against the expected circulating strains during the upcoming flu season. The vaccine composition varies due to strain variation, therefore circulation of the same subtype may require a different vaccine (Carrat & Flahault, 2007). Vaccination is highly recommended and considered to be the best means of preventing infection. For this reason, the vaccine is offered free of charge by the National Health Service to those considered to be at high risk, such as people aged over 65, young children, pregnant women, those with underlying health conditions and health care workers. However, the vaccine fails to provide long term immunity due to antigenic drifts (gradual accumulation of changes in the HA and NA genes) which permit the virus to re-infect a host. In February 2015, the high rate of evolution of the virus was exemplified by a mutation in the circulating H3N2 vaccine strain. This severely impacted the effectiveness of the vaccine, which was estimated to be only 19% effective during the 2014/2015 season (CDC, 2017; Chambers, Parkhouse, Ross, Alby, & Hensley, 2015). Furthermore, in certain individuals response to vaccination is poor, and can lead to serious complications of flu, highlighting the need for effective, novel antivirals which are unlikely to be affected by resistance mutations to limit virus transmission.

This chapter presents a review of influenza A virus biology, current antivirals and emerging inhibitors, and the rationale behind targeting evolutionary conserved binding sites of viral proteins. The work presented in subsequent chapters will focus on selected internal influenza proteins as antiviral targets, leading to the discovery of molecules that may serve as potential replication inhibitors.

1.2 Structure and infectious cycle of the influenza A virus

Influenza A belongs to the *Orthomyxoviridae* family of viruses which are characterised by a lipid bilayer envelope and a segmented, negative-sense single stranded RNA genome (figure 1a) (Cheung & Poon, 2007). This genome is ~13,600 base pairs in length and comprises eight different RNA segments that encode up to 17 proteins; several of which are produced from a single segment (Vasin et al., 2014). New proteins are continually being discovered, although not all of these proteins contribute to the infectious life cycle (figure 1b), nor are they found to be present in every virus subtype. Virus particles are pleomorphic, with the spherical forms having a diameter of ~100 nm and the filamentous forms often in excess of 300 nm (Bouvier & Palese, 2008).

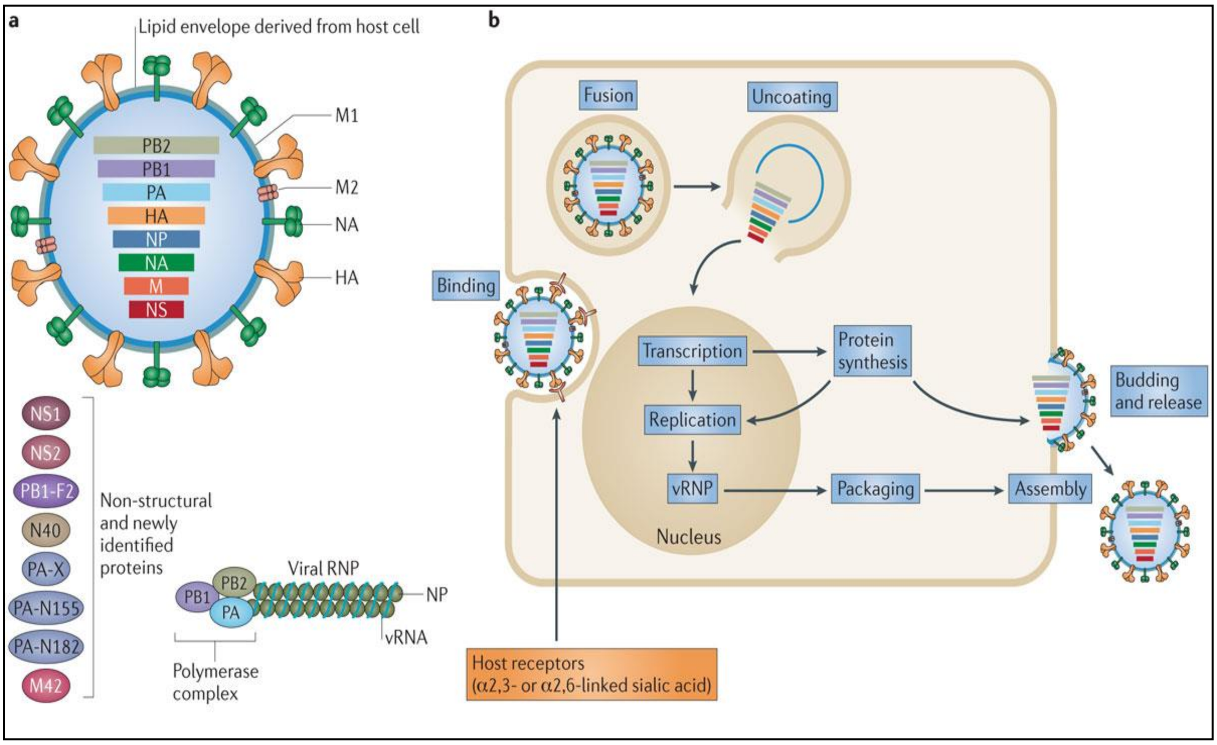


Figure 1. (a) Schematic structure and (b) life cycle of the influenza A virus. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Microbiology (Shi, Wu, Zhang, Qi, & Gao, 2014), copyright (2014) https://www.nature.com/nrmicro/.

The process of infection involves several stages, starting with the attachment protein haemaglutinin (HA) binding to sialic acid receptors on susceptible host cells such as epithelia of the upper respiratory tract. At this stage an antibody response against the HA antigen is usually activated in the host. Endocytosis occurs and the viral envelope merges into the membrane of the host cell allowing the nucleocapsid to become internalised in the cytoplasm (Cheung & Poon, 2007). The viral particle is held within endosomes where it is exposed to extremely acidic conditions. Prolonged acidification destroys the viral structure and triggers a large conformational change in the HA_2 subunit causing them to expose hydrophobic fusion-promoting regions. Upon viral and endosomal membrane fusion, the proton channel formed by the M2 protein is activated by the low pH and allows rapid transfer of hydrogen ions into the virus to facilitate uncoating of the capsid and dissociation of viral RNA (vRNA).

The vRNA genome is then released into the cytoplasm in the form of ribonucleoprotein (RNP) complexes, which consist of vRNA segments wrapped around nucleoprotein (NP) monomers, with the end of each segment bound by the polymerase complex (figure 1a, bottom). The NP also maintains the RNP structure and is involved in many other processes such as intracellular trafficking of the viral genome. From the cytoplasm, the vRNA segments translocate to the nucleus to be effectively transcribed into messenger RNA (mRNA) segments (Bouvier & Palese, 2008).

The influenza RNA dependent RNA polymerase (RdRp) composed of the three subunits: acidic polymerase (PA), polymerase basic protein 1 (PB1) and polymerase basic protein 2 (PB2) binds and cleaves capped host cell mRNA to use as a primer for viral mRNA synthesis. Positive-sense mRNA templates are produced for each segment and the viral mRNA is then translocated back into the cytoplasm in the form of RNP's where it is translated into virus proteins. Some RNA is also retained in the nucleus as complementary RNA, from which the newly synthesised negative-sense vRNA is formed to be used as a template for further genome replication, which occurs in a primer independent manner. The process of nuclear export is mediated by the NEP and M1 proteins (Akarsu et al., 2003).

The processed HA, NA and M2 surface proteins are transported and inserted in to the cell membrane, whilst the eight genomic segments are encapsidated by the NP, packaged and then enveloped using the host plasma membrane to form new virus particles. The virus finally buds from the infected cell and is released by sialidase enzyme activity of the NA approximately eight hours after infection (Bouvier & Palese, 2008; Couch, 1996). The influenza virus is notorious for replication errors, and if the cell is infected by two or more strains at the same time, random genetic re-assortments, which lead to genomic diversity without impairing constituent protein functions. The process of viral genome replication disrupts normal host cell physiology and biochemistry, and is targeted by the immune system causing the characteristic symptoms of infection. The proteins of the influenza A virus are summarised in table 2, many of which are involved in the infectious life cycle and can be considered as potential antiviral drug targets.

Table 2. Proteins encoded by the 8 RNA genome segments of the influenza A virus and their major functions (Bouvier & Palese, 2008; Vasin et al., 2014).

RNA segment	Protein encoded	Number of residues**	Function
1	PB2	759	Polymerase subunit: 5’ cap binding of host mRNA
2	PB1	757	Initiates RNA synthesis, RNA elongation
	PB1-F2*	~90	Pro-apoptotic virulence factor, interacts with PB1
	PB1-N40*	718	N-terminally truncated form of PB1, regulates PB1 and PB1-F2 expression
3	PA	716	Cleavage of capped host mRNA
	PA-N155*	568	Unknown (isoform of PA)
	PA-N182*	535	Unknown (isoform of PA)
	PA-X*	252	Modulates host response and virulence (isoform of PA)
4	HA	550	Major antigen; attachment to sialic acid residues on host cells
5	NP	498	Viral RNA binding, structural component of ribonucleoprotein complex, acts as adaptor between virus and host cells
6	NA	470	Cleaves bonds between HA and sialic acid facilitating virus release from cell surfaces
7	M1	252	Component of virion providing structural support, exports ribonucleoproteins
	M2	97	Internal proton channel; virus uncoating
	M42*	99	Can replace M2 in M2-null viruses
8	NS1	219	Suppresses host interferon based antiviral immune response
	NS2/NEP	121	Facilitates exit of viral RNA from the nucleus to cytoplasm
	NS3*	174	Isoform of NS1

*Recently discovered (since 2001), may not be present in all influenza A strains.
**The number of residues for each protein varies between strains.

1.3 Current influenza A antivirals and documented resistance

Although influenza infection is usually self-limiting, for high-risk patients or in case of highly pathogenic strains antiviral drug treatment is required. Two classes of antiviral drugs with specific activity targeting either the neuraminidase (NA) surface protein or the M2 proton channel are the first line of defence (figure 2). The NA inhibitors (Oseltamivir, Laninamivir, Peramivir and Zanamivir) block the enzymes active site which prevents cleavage of sialic acid residues on the surface of infected cells to stop the virus from spreading (Stiver, 2003). Whereas, the M2 inhibitors (adamantanes) obstruct the proton channel by interacting with the hydrophobic transmembrane domain of the M2 protein, thus preventing the entry of the viral genome into cells (Englund, 2002).

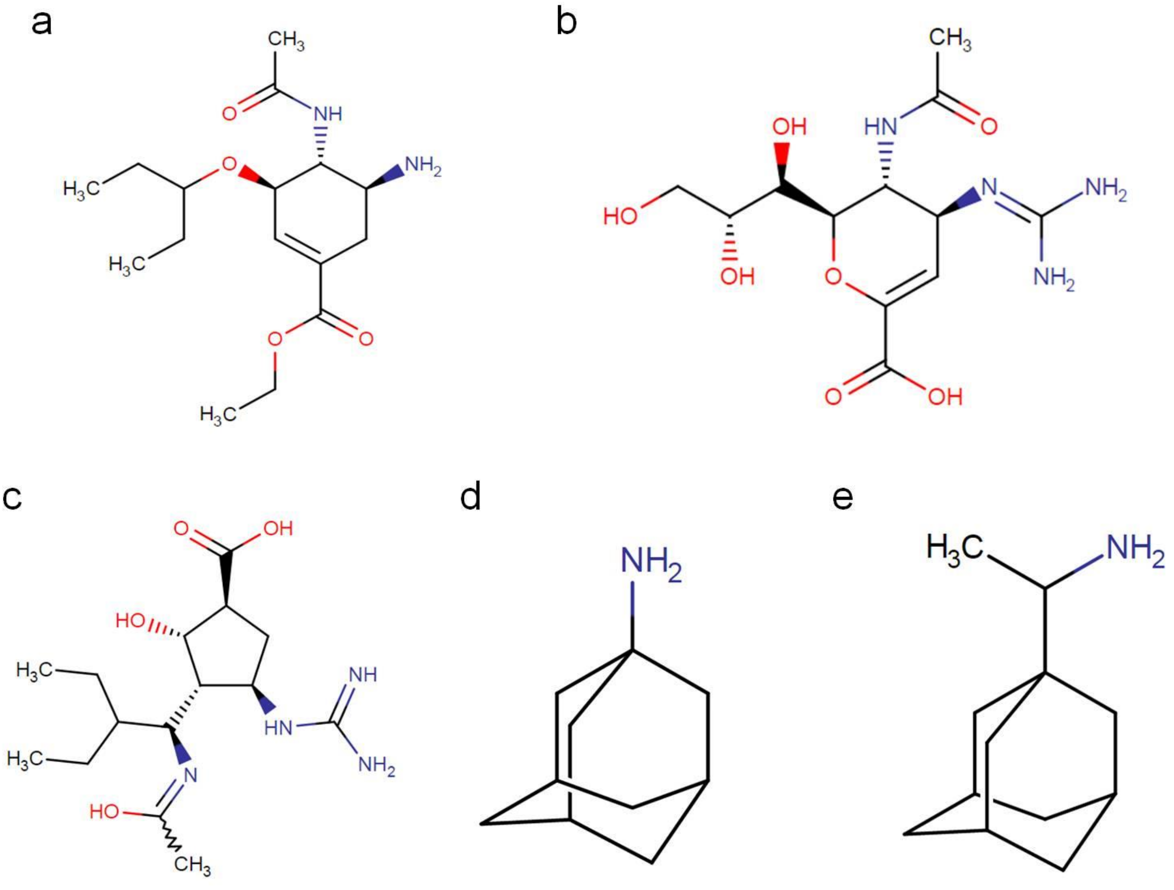


Figure 2. Chemical structures of NA inhibitors (a) Oseltamivir, (b) Zanamivir, (c) Peramivir, and M2 inhibitors (d) Amantadine and (e) Rimantadine. All structures were obtained from the DrugBank database.

However, rapid amino acid mutation rates in these proteins and increasing emergence of widespread and subtype dependent antiviral drug resistance are concerning. Such examples of resistance include seasonal influenza A H1 strains harbouring the histidine to tyrosine substitution at position 274 (H274Y), conveying high level Oseltamivir resistance through partial displacement of Oseltamivir out of the NA binding site, corresponding to antiviral treatment failure (Hurt, Holien, Parker, & Barr, 2009; Moscona, 2009). Similarly, the affinity of drug binding is reduced by the S31N substitution in the transmembrane domain of the M2 protein (Schnell & Chou, 2008) and due to extensive resistance, the adamantanes are no longer recommended for antiviral treatment (CDC, 2016). Several other amino acid point mutations in the NA and M2 proteins have also been found which are known to confer a resistant phenotype (Samson, Pizzorno, Abed, & Boivin, 2013; Wang et al., 2011). Furthermore, depending on the virus strain, these antiviral drugs may not work to the same extent due to differing levels in drug sensitivity.

1.4 Recent discoveries of influenza A inhibitors

Due to increasing reports of drug resistance, inhibitor molecules targeting proteins other than the M2 and NA are frequently being discovered through different approaches and experimentally evaluated by various methods (Naesens, Stevaert, & Vanderlinden, 2016). A number of NS1 protein antagonists such as NSC125044, JJ3297 and A22 were identified to reduce influenza replication using cell based assays, whilst other inhibitors have been found using functional biochemical assays (reviewed in Engel, 2013). Although the specific binding regions for many of these compounds to NS1 are unknown.

The inhibitor Nucleozin identified from a cell based screening study was found to target the nucleoprotein or the viral RNP complex (Amorim, Kao, & Digard, 2013; Kao et al., 2010) and has shown inhibitory effects on virus replication *in vitro*. The small molecule Naproxen also targeting the NP was initially identified by virtual screening, followed by molecular dynamics simulation analysis, and verified by *in vitro* antiviral tests to show reduced viral titres (Lejal et al., 2013). The compound RK424, identified from screening a library of compounds using cell based replication assays was also found to target the NP/RNP complex and reduced virus replication of several strains (Kakisaka et al., 2015).

The viral polymerase is another popular antiviral target of interest; the purine analog Favipiravir (T-705) is a drug that targets the RNA polymerase, and phase three clinical trials have been completed in Japan (Furuta et al., 2013). The novel small molecule inhibitor ASN2 which was discovered from a high-throughput cell screening assay also targets the polymerase (presumably the PB1 subunit) and can inhibit replication of the major pandemic subtypes (Ortigoza et al., 2012). More recently, a compound targeting the PB1-PB2 interface named PP7 has been discovered and showed inhibitory effects on virus replication against H1N1, H7N9 and H9N2 subtypes (Yuan et al., 2017). Overall, the number of potential influenza A replication inhibitors reported is steadily rising and there is increasing focus on investigating internal proteins as target sites for antiviral drugs (reviewed in Patel & Kukol, (2016b)). This is due to lower rates of evolution based on sequence and structure analysis (Warren, Wan, Conant, & Korkin, 2013).

1.5 Evolutionary conserved ligand binding sites as antiviral targets

Comparison of protein sequences enables similarities and differences at the level of individual amino acid residues to be analysed with the aim of inferring evolutionary relationships. Sequence relatedness often corresponds to the level of structural and functional conservation; therefore appreciating the evolution of influenza A protein sequences from all hosts and subtypes over time is essential for antiviral drug discovery. One common approach for this analysis is through multiple sequence alignment (Capra & Singh, 2007). Generally, residues displaying strict conservation are essential for correct protein folding, structure support and maintenance, as well as comprising binding interfaces and molecular recognition sites. Non-essential amino acids are usually more prone to mutation, with a minor influence on the viral fitness. However, a conservation study on the influenza acid polymerase protein has shown that residues classed as non-conserved may indeed be biologically important, and that functional residues are not always conserved (Wu et al., 2015).

Sequence conservation data may be incorporated with structural information available for proteins in public databases, in addition to computationally predicted models to identify and evaluate antiviral target sites as shown in this work. Previously, similar studies using this approach have identified novel binding sites for antivirals on the NS1 protein and nuclear export protein (Darapaneni, Prabhakar & Kukol, 2009), as well as the nucleoprotein (Kukol & Hughes, 2014). Building on this approach with the use of virtual screening and docking predictions, potential drug target sites and compounds can be further assessed. It is important that potential ligand binding sites are within or close to highly conserved amino acid regions as they are less likely to undergo genetic mutations which may render antiviral drugs targeting these sites ineffective (Kukol & Patel, 2014).

1.6 Aims & Objectives

To address the problem of influenza antiviral drug resistance, the reassortment of genomic segments from different host-type viruses, and the lack of effective drugs available, this research aimed to identify 1) highly conserved regions of selected internal influenza A viral proteins that overlap with predicted ligand binding sites and 2) potential inhibitor molecules that may bind to those sites.

These aims are addressed by drawing on the increased number of influenza virus genome sequences in comparison to earlier studies (Darapaneni, Prabhaker, & Kukol, 2009; Kukol & Hughes, 2014), as well as using an improved method of calculating sequence conservation and a more accurate method of predicting binding site locations. In addition, full-length protein structures are used with molecular dynamics simulations performed on the initial comparative models. Molecular docking based virtual screening is used to predict potential inhibitor molecules. Additional work presented in chapter six aimed to improve the accuracy of virtual screening and docking predictions by developing and evaluating a novel receptor-decoy binding site strategy.

2 GENERAL MATERIALS & METHODS

2.1 Materials

Local computer workstations with Linux and Windows operating systems were used for bioinformatics and protein structure analysis. Molecular dynamics (MD) simulations were performed with Gromacs versions 4.5.5, 4.6.5 or 2016 (Hess, Kutzner, van der Spoel, & Lindahl, 2008; Pronk et al., 2013) on the University of Hertfordshire High Performance Computer cluster (UH-HPC). Graphs were produced with the plotting tool GRACE, using .xvg format files as input. Virtual screening was performed with AutoDock version 4.2 (Morris & Huey, 2009; Morris et al., 1998) and AutoDock Vina version 1.1.1 (Trott & Olson, 2010) with the help of Raccoon graphical user interface (Forli et al., 2016) on the UH-HPC and AutoDock Tools version 1.5.6. Protein structures, simulation trajectories and docking results were visualised with the molecular graphics viewers Rasmol (Sayle & Milner-White, 1995), Visual Molecular Dynamics (VMD) (Humphrey, Dalke, & Schulten, 1996), The PyMol Molecular Graphics System, version 1.7.4.5 Edu, Schrodinger, LLC and AutoDock Tools (Morris & Huey, 2009). Molecular interactions between proteins and ligands were identified with LigPlot+ version 1.4.5 (Laskowski & Swindells, 2011).

2.2 Protein sequence analysis

All protein sequences were downloaded from the National Centre for Biotechnology Information (NCBI) Influenza Virus Resource database (Bao et al., 2008). The Cluster Database at High Identity with Tolerance (CD-HIT) Suite web server (Huang, Niu, Gao, Fu, & Li, 2010) was used to remove redundancy of sequences based on a specified similarity threshold. Multiple sequence alignments were performed with the Clustal Omega web server version 1.2.1 or 1.2.3 (Sievers et al., 2011). The default settings for all parameters remained unchanged whereby the mBed algorithm was used to generate the guide tree and the number of combined guide tree and Hidden Markov Model iterations remained at 0. The sequence alignment viewer and editor Jalview Version 2 (Waterhouse, Procter, Martin, Clamp, & Barton, 2009) and BioEdit version 7.2.3 (Hall, 1999) was used to analyse and edit the alignments.

2.3 Calculation of amino acid conservation

From the multiple sequence alignment output file, the amino acid conservation was calculated using the Jalview AACons Web server (Waterhouse et al., 2009). The Valdar scoring method was selected for the calculation as it showed to be an accurate approach to quantify evolutionary conservation, described in a review evaluating 18 different scoring methods. The parameters considered in the calculation include amino acid frequency and stereo-chemical diversity with a full substitution matrix, as well as normalising against redundancy to reduce the effect of bias sampling and penalizing gaps (Valdar, 2002). The calculation produces a numerical score for each position in the alignment ranging between 0 (low conservation) and 1.0 (high conservation). To map the degree of amino acid conservation onto the protein structures with a colour scale and identify conserved residues within binding sites, the original conservation scores obtained from the Jalview web server were re-scaled between zero and one hundred (0 being high conservation and 100 being low conservation) using the following formula:

$$New\ Score = (Vs - min) * (100/\ (1 - min))$$

Where *Vs* is the original Valdar score, *min* is the minimum conservation score out of the dataset and 1 is the maximum conservation score. The values for the temperature-factors in a PDB file were then replaced with the re-scaled conservation scores for each amino acid. Thus, allowing the protein structure to be coloured by conservation according to a temperature gradient from blue (high conservation/cold) to red (low conservation/hot).

2.4 Protein modelling and molecular dynamics (MD) simulations

Experimental structures of influenza proteins were obtained from the RCSB protein data bank (PDB) (www.rcsb.org). To account for incomplete or missing regions of structures, the Iterative Threading ASSembly Refinement (I-TASSER) server was used for 3D protein structure prediction (http://zhanglab.ccmb.med.umich.edu/I-TASSER/). The server uses both template based and template free (*ab-initio*) methods to generate models and consistently performs well in the biennial community wide Critical Assessment of protein Structure Prediction (CASP) experiments (Yang & Zhang, 2015; Zhang, 2008). Experimental and model coordinates were combined to generate full length models. Energy minimisation was performed to remove atomic clashes from the models and to prepare the system for MD simulations. MD simulations provide a dynamic view of protein structure over time and have been used to improve models from structure prediction methods (Kalia & Kukol, 2011). The accuracy of MD simulations is dependent on the force field which is the combination of mathematical equations and parameters used to relate the chemical structure to energy based on atomic positions (Guvench & MacKerell, 2008). In this work simulations over 100 ns were performed to reveal conformational changes of the unrestrained regions of the protein models.

2.4.1 Simulation analysis

Simulation trajectories were analysed by calculating the root mean square deviation (RMSD) of backbone heavy atoms with respect to the starting structure using the Gromacs function ‘g_rms’. The RMSD is a measure of overall structural similarity between structures ‘1’ and ‘0’ and is calculated as follows:

$$RMSD = \sqrt{\frac{1}{M}\sum_{i=1}^{N} m_i (r_i^0 - r_i^1)^2}$$

Where r_i = (x,y,z) is the vector of the coordinates, N the number of atoms, M the relative mass of the molecule and m_i the relative mass of each atom. A backbone RMSD of smaller than 0.2 nm is normally considered as small structural fluctuation; while an RMSD > 0.3 nm indicates a conformational change.

Cluster analysis was performed to find the most representative groups of protein structure conformations sampled during the simulations using the Gromacs function ‘g_cluster’. The RMSD cut-off distance between structures to determine cluster membership and the fraction of the trajectory frames to analyse were determined from the RMSD plots. The gromos method for clustering was used. The central structure is the structure with the smallest average RMSD from all other structures within that cluster. The group selected for least squares fitting was backbone and the group for output was selected as protein.

2.5 Prediction of binding hot spots

Computational solvent mapping to identify potential ligand binding hot spots was performed using the FTMap web server (http://ftmap.bu.edu/) (Hall, Kozakov, & Vajda, 2012) for all influenza A proteins selected. The server requires a structure in PDB format to be uploaded and uses 16 different small organic molecules of varying properties, (ethanol, isopropanol, isobutanol, acetone, acetaldehyde, dimethyl ether, cyclohexane, ethane, acetonitrile, urea, methylamine, phenol, benzaldehyde, benzene, acetamide, and N,N dimethylformamide) as probes which are docked onto the protein surface to locate favourable binding hot spots. These hot spots are ranked based on their average binding free energy, and low energy sites where several different probe clusters overlap (consensus) are considered as potential ligand binding sites. In addition to locating binding hot spots, functional groups from the probes that interact favourably with the protein can be identified (Brenke et al., 2009; Hall, Kozakov, & Vajda, 2012). The principle behind computational solvent mapping is shown in figure 3.

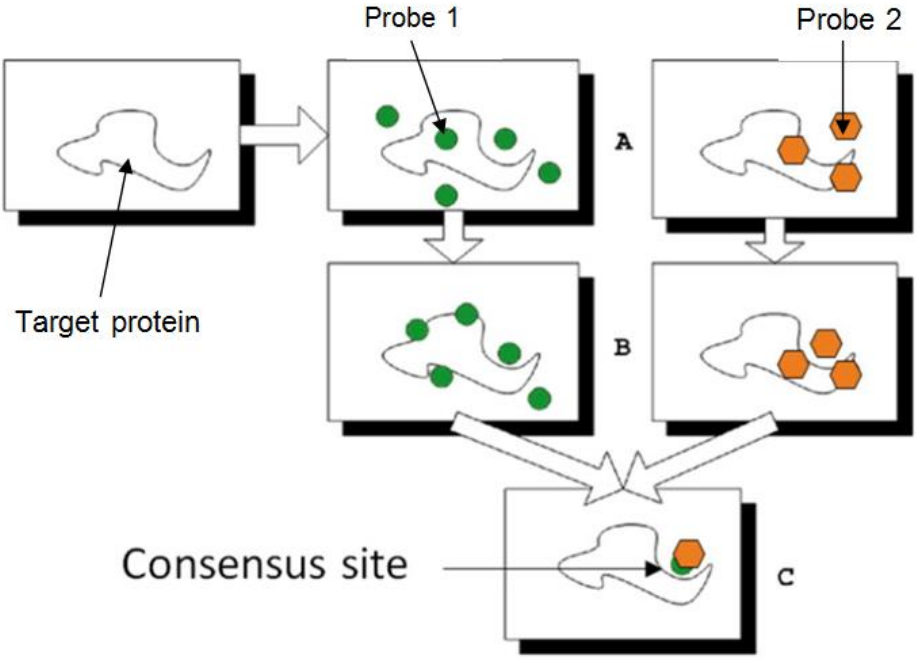


Figure 3. Principle of computational solvent mapping using two probes to identify ligand binding sites. The green circles and orange hexagons indicate a cluster for a molecular probe type. Where clusters overlap indicates the consensus site. Adapted by permission from Springer Nature: Springer, New York, NY, Computational Drug Discovery and Design, Hall et al., 2012, copyright (2012).

2.6 Virtual Screening

Virtual screening (VS) is a widely used technique in the field of medicinal chemistry to identify lead compounds from a chemically diverse library that can bind to a receptor. It is usually distinguished between ligand-based and receptor-based VS, with ligand-based VS requiring a number of known ligands to develop a pharmacophore model that is subsequently used to search a chemical library. In the present work the receptor based VS approach was followed that involves a process called molecular docking, which employs an algorithm to predict the best binding conformation of flexible small molecules to a specific protein receptor binding site (Kitchen, Decornez, Furr, & Bajorath, 2004). Three main stages are involved in the docking procedure: file preparation of the receptor, docking of the ligand through conformational and positional search and ligand scoring. The scoring and docking is often combined, as the scoring function is used to optimise the docking process, although external scoring functions can be applied afterwards. The scoring function used in the docking software evaluates the binding free energy of docking configurations to reflect the strength of protein-ligand binding. The results from screening large chemical libraries can then be ranked and the conformations with highest negative binding energy are considered to portray the most favourable binding mode, however, these results must be treated with caution (Chen, 2015). The rank list can be filtered and assessed to appropriately select compounds for experimental testing.

The factors that influence the outcome of a docking experiment include the biophysical interactions between protein and ligand on the one hand, namely, complementarity in size, shape, electrostatics, polarity and potential for hydrophobic, hydrogen bonding interactions between the ligand and receptor (Karthikeyan & Vyas, 2014), and on the other hand the exhaustiveness of the conformational search and accuracy of the scoring function. An evaluation of three docking algorithms that are commonly used for virtual screening showed that AutoDock Vina was the best single method, while ligand prediction can be improved by combining the rank lists of AutoDock 4 and AutoDock Vina (Kukol, 2011). This consensus approach was therefore used for the NS1 protein and the NEP.

2.6.1 Chemical compound library

The National Cancer Insitute (NCI) Plated 2007 chemical compound library was downloaded from the ZINC database (http://zinc.docking.org/). This consisted of 3D structures of all molecules at pH 6-8 in .mol2 format. The file containing the codes of molecules at 80% similarity cut-off was also obtained from the ZINC database that applied clustering based on the Tanimoto coefficient using ChemAxon default fingerprints. These molecules were in the simplified molecular input line entry system format (.smi) which describes chemical structures in a linear format that can be interpreted by the computer. A script (appendix 9.5) was used to extract the molecules within the 80% similarity cut-off level from the NCI library to give a chemically diverse library. The extracted molecules were written to an output file in .mol2 format for virtual screening using the command:

```
$ perl extract.pl result-80.smi nci_all.mol2 > output.mol2
```

The chemical compounds in .mol2 format were split into individual files and saved in the PDBQT format using the AutoDock screening preparation tool Raccoon to give a total of 52,172 compounds. The PDBQT format is an extension of the PDB format with the addition of atom charges and types added into the file.

2.6.2 AutoDock 4

AutoDock 4 is freely available software which performs docking based on a grid map method to describe the receptor protein and the Lamarckian Genetic Algorithm to iteratively select the lowest binding energy conformation of a ligand (Morris et al., 2009). The AutoDock Tools software package (Morris & Huey, 2009) was used to define the 3D space for docking with a grid spacing of 0.375 Å to calculate box dimensions. A grid parameter file (.gpf) specifying the size and location of the docking grid and a docking parameter file (.dpf) specifying the input parameters for the AutoDock calculation must be written to perform the docking (appendix 9.6 and 9.7). For all protein targets the number of energy evaluations was set to 350,000 and the population size was 150. The screening preparation tool Raccoon was used to prepare files; the interface consists of five tabs: Ligand(s), Receptor(s), Maps, Docking and VS Generation, which required input of the ligands and receptor files in PDBQT format, the .gpf to produce the grid maps for each atom type in the ligands, and the .dpf. The script used to perform the screening is shown in appendix 9.8.

2.6.3 AutoDock Vina

AutoDock Vina is also a freely available program for virtual screening and molecular docking. Although it has a similar name to AutoDock 4 it is a different program with higher speed and accuracy due to a different docking algorithm and scoring function implemented to predict binding modes (Trott & Olson, 2010). To perform the screening with AutoDock Vina, a configuration file was written which specified the input files and the docking box size which remained the same as that used with AutoDock 4. The script used to perform the screening is shown in appendix 9.9.

To analyse the screening results, scripts were used to produce a rank list of binding affinities in kcal/mol from lowest to highest for each protein-ligand interaction from the AutoDock Vina and AutoDock 4 output files, and to combine the AutoDock 4 and AutoDock Vina rank lists together (appendix 9.10).

3 THE NON-STRUCTURAL PROTEIN 1 (NS1)

3.1 INTRODUCTION

The influenza A non-structural protein 1 (figure 4) is a multifunctional protein encoded by the unspliced mRNA of genome segment eight. It is approximately 26 kDa in weight and its most recognised function is counteracting the interferon (IFN) based host immune response via two molecular mechanisms: inhibiting activation of interferon regulatory factor 3 (IRF3) and therefore reducing IFN gene expression (Mibayashi et al., 2007; Talon et al., 2000), and inhibiting IFN pre-mRNA processing through binding CPSF30; although these functions are reported to be strain specific (Krug, 2015). Other important functions include binding poly A tails of cellular mRNA keeping them in the nucleus, and regulating viral RNA replication (Hale, Randall, Ortin, & Jackson, 2008). Previously, the NS1 protein was thought to be non-structural, however, it has recently been identified at low levels within nascent virions and can therefore be considered a structural component of the virion (Hutchinson et al., 2014). The NS1 structure consists of an N-terminal RNA binding domain (RBD), and a C-terminal effector domain (ED) which is joined by a linker region (LR) (Lin, Lan, & Zhang, 2007), as shown in figure 4. A review by Hale (2014) states that NS1 can be considered a structurally dynamic protein that forms various conformational states.

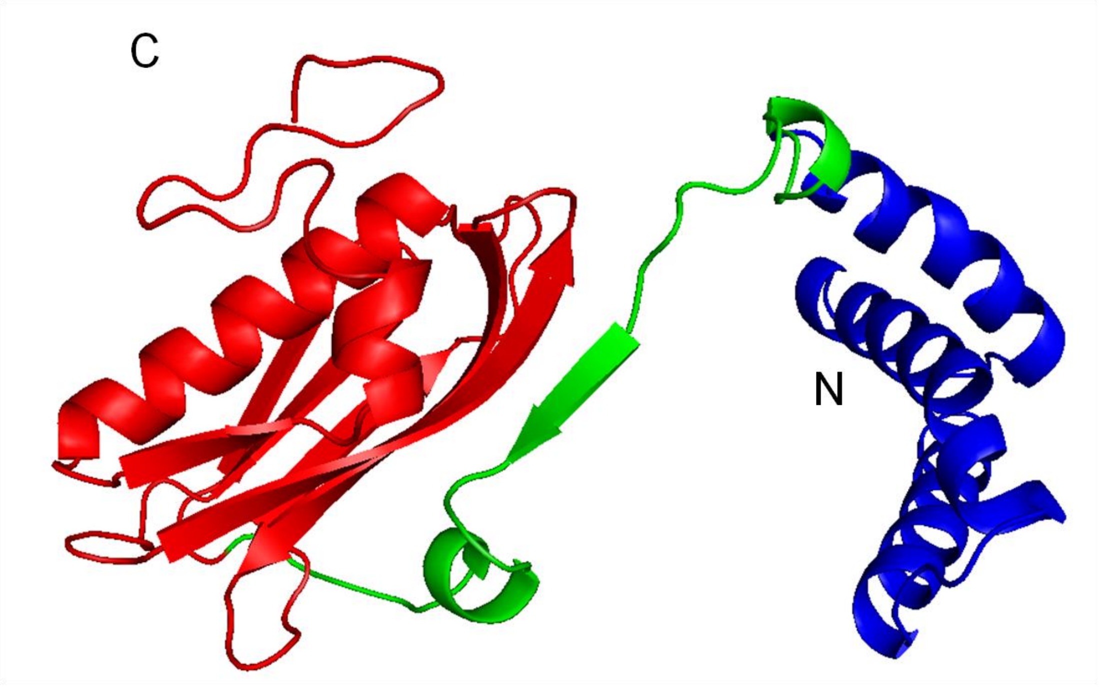


Figure 4. Cartoon representation of an NS1 protein model based on a H5N1 strain. The protein backbone is coloured showing the RBD in blue, LR in green and ED in red.

3.1.1 Structure and function of the RNA binding domain (RBD)

The RBD is made up of residues 1-73 which form three α-helices consisting of residues 4-24 (helix 1), 31-50 (helix 2) and 54-70 (helix 3) (Lin et al., 2007). The RBD facilitates the formation of NS1 homodimers in solution to recognise viral and cellular RNA molecules by interlocking three helices from one monomer with three helices from another monomer. It is reported that highly conserved basic residues such as Arg38 in helix two specifically interact through hydrogen bonding with the double stranded RNA backbone (Cheng, Wong, & Yuan, 2009). The linker region (LR) immediately following the RBD is made up of residues 74-88. The LR allows conformational changes of the RBD and ED and a recent crystallography study based on a H6N6 NS1 mutant revealed that a shortened LR, as well as residue identity at this position restricts the ability of the ED to adopt different positions relative to the RBD (Carrillo et al., 2014).

3.1.2 Structure and function of the effector domain (ED)

The ED is made up of residues ~88-230 which form three α-helices and seven β-strands (Bornholdt & Prasad, 2008). In this domain, six β–strands form an antiparallel twisted β-sheet around a long central α-helix and the structure is maintained through hydrophobic interactions between the twisted β-sheet and the α-helix (Lin et al., 2007). The ED inhibits post-transcriptional processing of cellular pre-mRNA by binding and inhibiting two cellular proteins: cleavage and polyadenylation specificity factor 30 (CPSF30), and poly (A) binding protein II (PABII). Residues between positions 144-188 bind two zinc finger regions of the CPSF30 protein which prevents the 3' end processing of cellular pre-mRNA. This results in suppression of host antiviral mRNA expression, such as IFN-β, reducing the host immune response (Twu, Noah, Rao, Kuo, & Krug, 2006). The 215-230 region binds PABII, also resulting in nuclear accumulation of unprocessed 3' poly A tail pre-mRNAs (Chen, Li, & Krug, 1999; Tu et al., 2011), inhibiting export of host mRNA and cellular protein synthesis.

The ED also binds protein kinase R (PKR) at the 123-127 region to deactivate the PKR pathway that would normally inhibit cellular and viral protein synthesis and consequently virus replication (Min, Li, Sen, & Krug, 2007). The C-terminal tail constitutes the last few residues of the ED and similar to the LR, varies in length depending on the strain (Carrillo et al., 2014). The changes in the length of the NS1 protein may be due to mutations associated with stop codons. Overall, as the NS1 ED contains several regions involved in protein binding that enhance virus survival; targeting these areas with drug-like compounds may inhibit key functions or protein interactions to influence virus replication. A study on NS1 conservation and binding site analysis has previously been done (Darapaneni et al., 2009), however, since 2009 the number of available NS1 sequences has increased more than three times and alternative computational methods with improved accuracy for calculating conservation (Valdar, 2002) and binding site prediction have been identified that are employed in this project. Additionally, through virtual screening, potential replication inhibitors are predicted and subsequently evaluated in cell-culture.

3.2 COMPUTATIONAL METHODS

3.2.1 NS1 sequence analysis

The sequence analysis was performed according to general methods section 2.2. Briefly, full length NS1 sequences were chosen from all hosts, regions and subtypes until 2014. Identical sequences were removed. The CD-HIT web server (Huang, Niu, Gao, Fu & Li, 2010) was used to remove redundancy of the sequences at a similarity threshold of 98%. Global pairwise sequence alignment between the NS1 H5N1 sequence A/Vietnam/1203/2004 used for protein modelling and the H1N1 sequence A/England/529/2013 of the strain used in experimental assays (gene sequence provided by Dr Angie Lackenby) was done using the EMBOSS Needle tool version 6.6.0 (http://www.ebi.ac.uk/Tools/psa/emboss_needle/) with the default settings. For calculation of amino acid conservation see general methods section 2.3. The minimum score used for re-scaling was 0.480. Scores for the extra five residues of the linker region and the extended C-terminal tail were not included during re-scaling.

3.2.2 Protein modelling

The full length H5N1 NS1 sequence (A/Vietnam/1203/2004) was submitted to the I-TASSER modelling server (Zhang, 2008) to identify a suitable template covering the entire protein sequence, and to build a structure that accounted for missing residues: 1-4, 75-79 and 198-215 in the PDB structure 3F5T. The RMSD between the I-TASSER model and PDB structure was calculated to indicate the level of similarity between the two sets of atom coordinates. The final NS1 model was generated by copying the structural information of residues 1-5, 74-80, 197-215 from the I-TASSER model into the PDB file of the experimental structure using a text editor. Missing side chains of Gln63 and Ile68 were reconstructed with Swiss-PDB Viewer version 4.1 (Guex & Peitsch, 1997).

3.2.3 Molecular dynamics simulations

The starting structure in PDB format was converted to a Gromacs file and molecular topology file with the program pdb2gmx. The AMBER99SB – ILDN force field was selected as it demonstrated high accuracy in a systematic evaluation of eleven molecular dynamics force fields using Gromacs (Beauchamp, Lin, Das, & Pande, 2012) along with the Transferable Intermolecular Potential 3 Point (TIP3P) water model. The protein was solvated in water in a cubic box with periodic boundary conditions to avoid edge effects at the boundary of the system. The charge of the chemical system was neutralised with four sodium ions and additional NaCl at 100 mM concentration. The following commands were used (‘$’ identifies the command prompt):

```
$ pdb2gmx -f ns1.pdb -o ns1.gro -p ns1.top -ter –ignh

$ editconf -f ns1.gro -d 0.75 -o box.gro -bt cubic

$ genbox -cp box.gro -cs spc216.gro -p ns1.top –o water.gro

$ grompp -f ions.mdp –p ns1.top –c water.gro -o ions.tpr

$ genion -s ions.tpr -neutral -conc 0.1 –p ns1.top -o ions.gro
```

Energy minimisation with 500 steps of the steepest descent algorithm followed by 200 ps MD simulation with position restraints on the protein non-hydrogen atoms was performed to equilibrate the system. The following commands were used:

```
$ grompp -f em.mdp -p ns1.top -c ions.gro -o em.tpr

$ mdrun -s em.tpr -c after_em.gro

$ grompp -f posres.mdp -p ns1.top -c after_em.gro -o posres.tpr

$ mdrun -s posres.tpr -c after_posresMD.gro -v >& posres.log &
```

For the production run, three replicate partially position restrained simulations over 100 ns (50 million steps) were performed. The position restraints were introduced to maintain the experimentally known structure, except the unknown C-terminus residues 196 to 215; therefore position restraints were gradually released as shown in table 3. The temperature of the system was set to remain constant at 300 K using the velocity rescaling thermostat with a coupling constant of 0.5 ps and pressure coupling was set at 1.0 atm with the Parinello Rahman barostat. The Particle Mesh Ewald (PME) method for calculating electrostatic interactions was used and the simulation was performed at a time step of 2 fs. Coordinates were written to the output trajectory file every ten thousand steps. The implementation of position restraints was verified by analysing root mean square fluctuation (RMSF) for each residue using the Gromacs function ‘g_rmsf’.

Table 3. Implementation of position restraint forces in the x, y, z axis on the non-hydrogen atoms of the NS1 protein. A force of 1000 indicates full restraint on an atom and 0 indicates full motion.

Atoms	Residue	fx	fy	fz
1-3081	1-195	1000	1000	1000
3082-3101	196	500	500	500
3102-3111	197	100	100	100
3112-3159	198-199	1	1	1
3160 onwards	200-215	0	0	0

3.2.4 Prediction of binding hot spots

Following cluster analysis, a central NS1 structure was selected for prediction of binding hot spots using FTMap. See general methods section 2.5.

3.2.5 Virtual screening

The screening was performed according to general methods section 2.6. The grid box dimensions for docking against the target site were set using AutoDock Tools as follows:

```
center_x = -6.147
center_y = 14.692
center_z = -21.069
size_x = 18.75
size_y = 16.5
size_z = 18.75
```

The receptor-decoy method described in chapter 6 was applied for the NS1 protein as initial results from the evaluation showed that this method was able to improve virtual screening predications. The grid box dimensions for the decoy site docking were set as follows:

```
center_x = 4.217
center_y = 5.529
center_z = -22.353
size_x = 15
size_y = 13.5
size_z = 18.75
```

Molecules in the top 15% of the decoy site rank list were used to generate the adjusted binding site rank list.

3.2.6 Potential drug target identification – PharmMapper

Following experimental work (section 3.3), the PharmMapper web server (http://59.78.96.61/pharmmapper/) (Liu et al., 2010) was used to identify potential protein drug targets given a small molecule. The mol2 file for compound D was submitted to the server using the default settings with all target sets selected. The server uses a pharmacophore mapping approach to screen a database (PharmTargetDB) which consists of known pharmacophore models (a collection of molecular features that enable interactions between a receptor and ligand) to find the best pose and 'fit score' of a query molecule to a protein target (Liu et al., 2010).

3.3 EXPERIMENTAL METHODS

Based on results from the consensus virtual screening predictions, selected compounds were ordered from the Drug Synthesis and Chemistry Branch, Developmental Therapeutics Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute (NCI) (http://dtp.cancer.gov) for antiviral testing. These compounds are supplied without charge, except postage, to the scientific community. The following experiments involving virus culture were performed at the Respiratory Virus Unit, Public Health England, Colindale, according to standard operating procedures.

3.3.1 Influenza virus titration by plaque assay

Plaque assays are used to determine the number of infectious viral particles in a sample. Therefore this assay was performed to determine the optimal virus dilution to infect the cell monolayer and produce a visible number of plaques for further antiviral experiments. The virus strain used for infection of cells was A/England/529/2013 (H1N1), which is a contemporary strain, sensitive to the neuraminidase inhibitor Oseltamivir carboxylate.

3.3.1.1 Preparation of cells

1 ml Madine-darby Canine Kidney-SIAT (MDCK-SIAT) cells in minimum essential media (MEM) with 10% Fetal Calf Serum were plated out into each well of a twelve well sterile flat bottomed tissue culture plate in triplicate. Plates were incubated for three days at 37°C with 5% CO_2 to produce a confluent cell monolayer.

3.3.1.2 Preparation of virus dilutions and inoculating cells

A ten-fold dilution series from 10^{-1} to 10^{-7} was prepared by adding 100 µl of virus suspension to 900 µl of virus transport medium (VTM). 100 µl of this dilution was then transferred to 900 µl of VTM and the dilution series was continued in this manner up to 10^{-7}. Growth medium was aspirated out of each well and the cell monolayer was washed twice with warmed, sterile phosphate buffer saline (PBS). To each well, 100 µl of the corresponding virus dilution was added; each dilution had duplicate wells. The virus was left to adsorb at room temperature for one hour. Plates were agitated gently every fifteen minutes to ensure an even distribution of the virus inoculum.

3.3.1.3 Preparation of agar overlay

10 ml aliquots of 2% MP Biomedicals Agarose were placed in a boiling water bath for fifteen minutes. The molten agar was then placed in a water bath at 45°C to cool before use. 1 ml of Gentamycin was added to 500 ml 2x Dulbecco's Modified Eagle's Media (DMEM) with Earles salts, 25 mM HEPES and Glutamax. 175 µl tolylsulfonyl

phenylalanine chloromethyl ketone (TPCK)-trypsin was added to 70 ml 2x DMEM to give a concentration of 1.25 µg/ml in the final agar overlay.

The virus inoculum was aspirated from the highest dilution to the lowest. 10 ml of the prepared warmed medium was added to 10 ml molten agarose and mixed. 1 ml of the agar/MEM mix was added to each well of the plate. The plates were left to stand at room temperature for fifteen minutes to cool and set, and then incubated at 37˚C and 5% CO_2 for 72 hours.

3.3.1.4 Fixing and staining cells

1 ml of 5% gluteraldehyde solution was added to each well and left at room temperature for a minimum of two hours to allow penetration through the agar. The excess gluteraldehyde was poured off and the agar was removed from each well by holding the plates in running water. 1 ml of 5% carbol fuchsin was added to each well to stain the cells. The stain was left for 30 minutes after which the excess carbol fuchsin was poured off. The plates were washed in running water to remove excess stain and blotted dry.

3.3.1.5 Calculation of plaque forming units (PFU)

The number of PFU/ml was calculated as follows:

Mean plaque number x virus dilution x 10

3.3.2 Influenza plaque reduction assay

A number of laboratory methods to assess the susceptibility of influenza to antiviral drugs have been developed (Sidwell & Smee, 2000). The plaque reduction assay is considered to be the gold standard method for assessing the susceptibility of influenza to antiviral compounds (Hayden, Cote, & Douglas, 1980; Zambon, 1998). The assay is based on assessing changes in plaque morphology (size and/or number) in the presence of an antiviral compound. The final result is usually expressed as a fold reduction in susceptibility to a drug compared to a reference virus control. However, it is not always possible to calculate an IC_{50} (50% inhibitory concentration) value using this assay.

From the results of the standard plaque assay, the virus dilution used for infection of the cell monolayer was $x10^{-4}$. The standard plaque assay using this dilution (with no drug) was also performed alongside each plaque reduction assay. 4.5 mg of compounds in table 6 were dissolved in Dimethyl Sulfoxide (DMSO) with heating to an initial stock concentration of 10 mM. 4.0 mg of compound H was dissolved in 100% ethyl acetate to an initial stock concentration of 10 mM.

3.3.2.1 Preparation and inoculation of cells

Cells were prepared as described in section 3.3.1.1. For the inoculation, 100 µl of virus was added to eleven wells, and 100 µl of VTM was added to the twelfth well (cell control). The virus was left to adsorb at room temperature for one hour. Plates were agitated gently every fifteen minutes to ensure an even distribution of the virus inoculum. An extra plate with six wells for virus controls, and six wells of media containing 1% DMSO as a cell control (no virus) was also prepared. A 1% DMSO concentration in 2x DMEM was tested as a cell control to assess the effect of DMSO with a 100 µM drug concentration.

3.3.2.2 Preparation of overlay and drug titration

Agar was prepared as described in section 3.3.1.3. 1 ml of Gentamycin was added to 500 ml 2x Dulbecco's Modified Eagle's Media (DMEM) with Earles salts, 25 mM HEPES and Glutamax. 500 µl DMSO was added to the media to give a final DMSO concentration of 0.1% in the assay. Starting with an initial concentration of 200 µM compound in warmed media, a ten-fold titration described in table 4 was performed. A ten-fold titration was also performed with Oseltamivir carboxylate for use as a positive control in the assay. 3.75 µl trypsin was added to each drug dilution to give a concentration of 1.25 µg/ml in the final agar overlay.

Table 4. Preparation of ten-fold drug dilutions for the plaque reduction assay.

Step	**Dilution series**	**Concentration (µM)**	**In assay concentration (µM)**
1	30 µl of 10 mM compound + 1470 µl 2x media	200	100
2	150 µl of step 1 + 1350 µl 2x media	20	10
3	150 µl of step 2 + 1350 µl 2x media	2	1
4	150 µl of step 3 + 1350 µl 2x media	0.2	0.1
5	150 µl of step 4 + 1350 µl 2x media	0.02	0.01

3.3.2.3 Removal of inoculum and addition of overlay

The inoculum was removed from all wells except the virus control and cell control wells. 1.5 ml of molten agarose was added to 1.5 ml of the appropriate media containing compound and then added to the corresponding wells in duplicate (figure 5). The inoculum was then removed from the virus control and cell control wells and 1.5 ml 2x DMEM/agar mix was added. The plates were left to stand at room temperature for fifteen minutes to cool and set, and then incubated at 37˚C and 5% CO_2 for 72 hours.

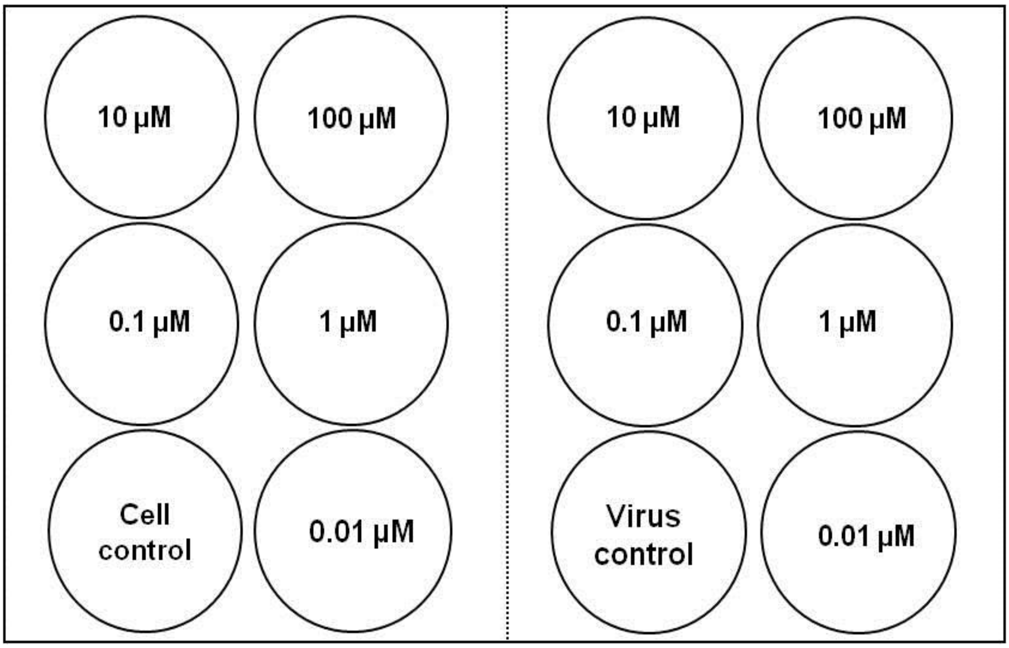


Figure 5. Layout of a twelve well plate overlaid with a ten-fold drug dilution series in duplicate for assessing plaque reduction.

3.3.2.4 Fixing and staining cells

Cells were fixed and stained as described in section 3.3.1.4.

3.3.3 Drug Affinity Responsive Target Stability (DARTS)

To determine if a compound had bound to the NS1 protein, the DARTS method was performed. This method is based on the concept that when a small molecule compound binds to a protein, the target protein structure becomes more stable and more resistant to degradation by proteases than the protein without the bound compound (figure 6). The extent of proteolysis can then be analysed by Western Blotting. The advantage of this method is that it can be performed using complex protein mixtures, without requiring purified proteins (Lomenick et al., 2009; Lomenick, Jung, Wohlschlegel, & Huang, 2011).

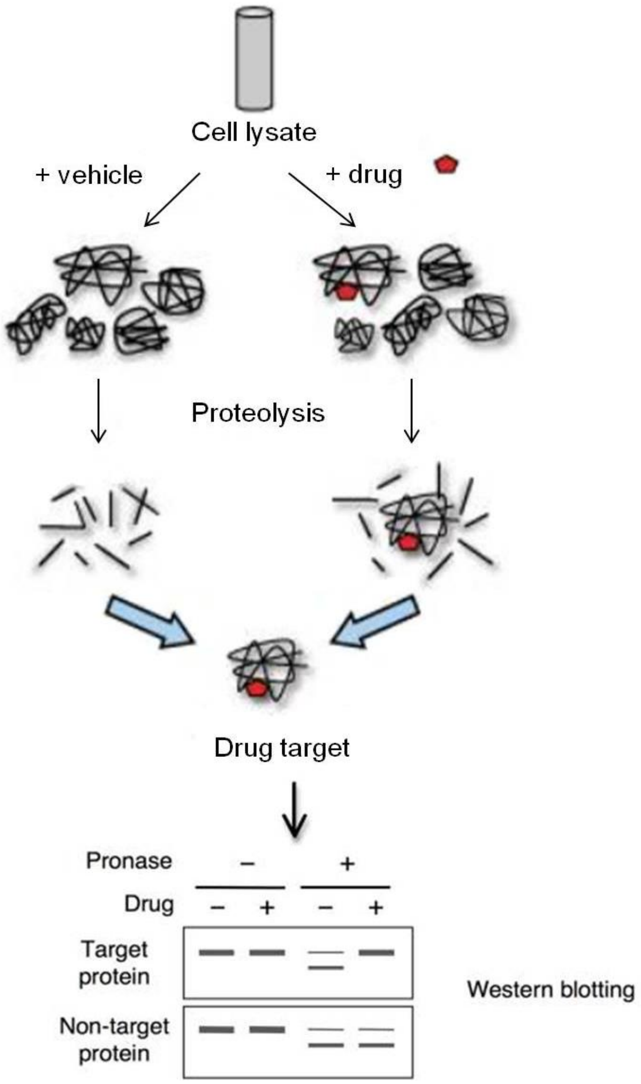


Figure 6. Diagram of the DARTS method for drug target identification. Figure adapted from (Lomenick et al., 2009, 2011). Reproduced with permission of the Licensor through PLSclear from Current Protocols in Chemical Biology, John Wiley & Sons Limited, copyright (2011).

3.3.3.1 Collecting and lysing cells

1 ml MDCK-SIAT cells were seeded in a twelve well tissue culture plate and incubated for three days at 37°C with 5% CO_2 to produce a confluent cell monolayer. The growth media was aspirated and washed twice with phosphate buffered saline (PBS). 100 µl of virus (A/Eng/529/2013) was inoculated into each well. The virus was left to adsorb at room temperature for one hour. Plates were agitated gently every fifteen minutes to ensure an even distribution of the virus inoculum. The inoculum was removed and 1 ml of serum free DMEM with 12.5 µg/ml of TPCK trypsin was added to each well. The cells were incubated at 37°C with 5% CO_2 for 24 hours.

To confirm viral replication, the virus titer from 50 µl of the supernatant of each well was determined by haemagglutination assay. The supernatants were then removed from each well and cell monolayers were washed once with PBS. 100 µl of 1x NuPAge lithium dodecyl sulphate (LDS) sample buffer (Thermo Fisher Scientific) was added to each well and left for ten minutes at room temperature. The cells were scraped off each well into the sample buffer. The buffer/cell mixture was heated at 95°C for five minutes and microfuged for 20 seconds.

3.3.3.2 Protein-small molecule incubation

3 µl dithiothreitol (DTT) reducing agent was added to 600 µl of the LDS sample buffer/cell lysate and heated at 80˚C for five minutes. The lysate was split into two samples (297 µl each). 3 µl DMSO was added to one sample and 3 µl small molecule (compound D) was added at 1 mM concentration to give a final concentration of 10 µM for a 1:100 dilution. The samples were mixed immediately by gentle flicking several times and microfuged briefly. The samples were incubated at room temperature for one hour.

3.3.3.3 Proteolysis

10 mg/ml Pronase stock solution was diluted to 1.25 mg/ml by mixing 12.5 µl Pronase with 87.5 µl cold 1x TNC buffer. This served as the 1:100 Pronase stock solution. This was diluted by mixing with 1x TNC buffer to create 1:300, 1:1000, 1:3000 and 1:10000 Pronase stock solutions.

Five 50 µl aliquots from both protein samples were prepared and the remaining 50 µl was saved as a non-digested control sample. 2 µl 1:100 Pronase solution was added to one aliquot of compound treated sample, mixed and incubated at room temperature. Exactly one minute after starting the first digestion, 2 µl 1:100 Pronase solution was added to one aliquot of the DMSO sample, mixed and incubated at room temperature. The eight remaining aliquots were digested at one minute intervals in this manner. After 30 minutes, the digestion of the first aliquot was stopped by adding 3 µl cold 20x Protease inhibitor solution, mixing and placing on ice. The remaining digestions were also stopped in the order they were started in one minute intervals.

3.3.3.4 Sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE)

SDS-PAGE is a common method for separation of protein mixtures by their molecular mass. 10 µl of the non-digested sample and each digested sample (compound treated and DMSO treated) were loaded and separated on a 15% polyacrylamide gel. 12 µl of biotinylated protein ladder (Cell Signalling Technology, #7727) was loaded alongside the samples as a molecular weight reference. The electrophoresis was run at 200V (constant voltage) for 40 minutes.

3.3.3.5 Western Blotting

Following SDS-PAGE, the gel was rinsed in blotting buffer for 15 minutes to remove salts and detergents. A polyvinylidene difluoride (PVDF) membrane was soaked in methanol for 15 seconds, then water for two minutes, followed by blotting buffer for 15 minutes. The gel was transferred onto the PVDF membrane by wet blotting and the transfer was carried out at 200mA for 1 hour 45 minutes. The PVDF membrane was washed in 20 ml blocking solution (2% Tween 80 in PBS with 5% milk powder) for forty five minutes with shaking at room temperature. The membrane was then incubated in 15 ml of this blocking solution with addition of the influenza A virus NS1 primary polyclonal antibody (Thermo Fisher Scientifc, #PA5-32243) at a 1:2000 dilution at 4˚C overnight with shaking.

The membrane was washed four times every five minutes in washing solution (PBS-Tween x1) with shaking to remove traces of any free primary antibody. The membrane was then incubated in 10 ml of washing solution with addition of the secondary antibody (Goat anti-rabbit IgG HRP conjugated, Thermo Fisher Scientifc) at a 1:1000 dilution, and a HRP conjugated anti-biotin antibody at a 1:1000 dilution to detect the biotinylated protein ladder for one hour at room temperature with shaking. The membrane was washed again four times every five minutes in washing solution with shaking to remove unbound antibodies.

3.3.3.6 Enhanced Chemiluminescence (ECL) detection

Equal volumes of reagent 1 and reagent 2 (Pierce^{TM} ECL Western Blotting Substrate, Thermo Fisher Scientific) were mixed together, added to the membrane and left to incubate for five minutes at room temperature. The protein blot was captured using the myECL Imager (Thermo Fisher Scientific).

3.4 RESULTS

3.4.1 NS1 amino acid conservation

8426 NS1 protein sequences mainly from human and avian hosts from all geographic regions were initially obtained from the NCBI Influenza Virus Resource database (Bao et al., 2008). Applying a similarity threshold of 98%, the redundancy was reduced with CD-HIT. Sequences with non-standard residues were removed resulting in 1416 remaining sequences. In this work, the H5N1 A/Vietnam/1203/2004 sequence numbering convention has been used where the five amino acid deletion is not included. This deletion found in several sequences was accounted for by the insertion of a gap region in the alignment.

The Valdar conservation scores ranged from 0.480 (lowest) at position 215 to 1.0 (highest) with Pro31, Asp115 and Ala127 being 100% conserved. The alignment profile showed that the majority of NS1 sequences had a high level of conservation in certain areas of the RNA binding domain and effector domain. This includes regions from Ile112-Lys121, and Leu142-Ser160, which both form beta sheets and loops on the effector domain. Regions of intermediate or low conservation (<0.850) include positions 21, 22, 25, 26, 60, 48, 73-79 (linker region), 107, 166, 201 and 212 onwards (C-terminal tail). Amino acid substitutions observed for low conservation residues include E, G, D, A, V, I and S at position 60 and G, A, T, D, N, Y, S, R, K and I at position 166. The Valdar conservation scores for each amino acid position are presented in table 18 (appendix 9.1). The re-scaled scores were mapped to the NS1 structure for display purposes as shown in figure 7.

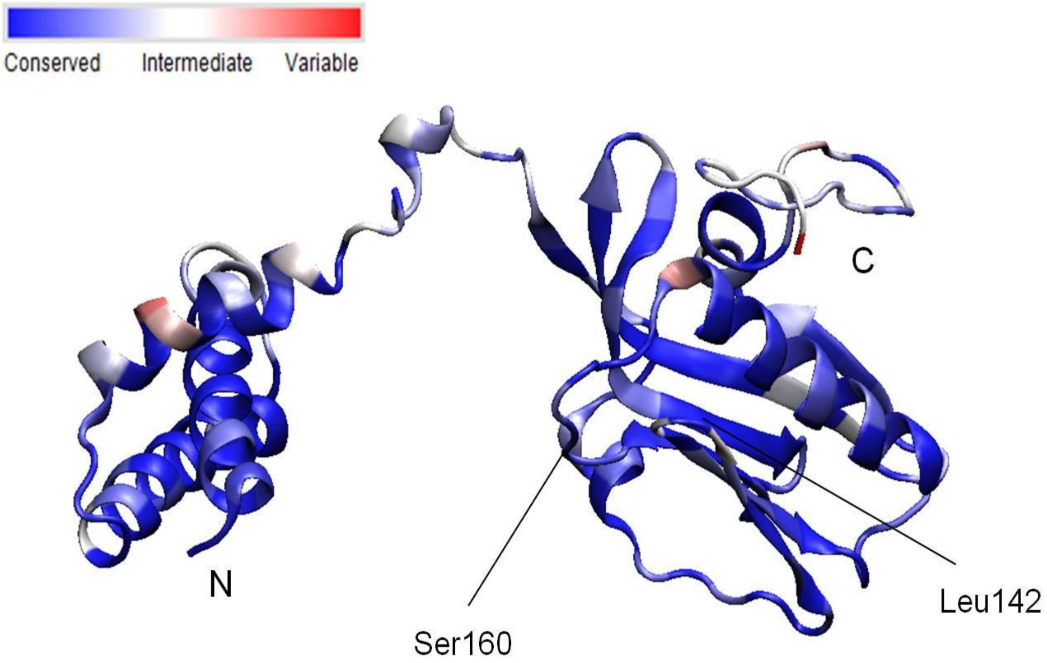


Figure 7. NS1 amino acid conservation mapped onto the protein structure. Blue regions indicate high conservation and red regions indicate low conservation (see colour scale).

3.4.2 Protein modelling and MD simulations

The I-TASSER server generated a full length NS1 model based on the H5N1 input sequence. Regions of the structure that were not sufficiently similar to any template were modelled *ab-initio* by the I-TASSER method (Zhang, 2008). The I-TASSER model and PDB experimental structure were somewhat different; in particular the side chain locations. The RMSD between backbone atoms was 0.87 Å, between all atoms 1.70 Å, and between side chains: 2.26 Å. Therefore the final NS1 model was produced by copying the structural information of residues 1-5, 74-80 and 197-215 from the I-TASSER model into the PDB file of the experimental structure.

From three repeats of 100 ns simulation trajectories and cluster analysis of the whole protein, it was observed that the conformation of the C-terminus (Trp198-Arg215) changed from the initial I-TASSER model; the experimentally known parts of the structure were kept fixed in space during the simulation. The first simulation produced three clusters of structures with cluster one being dominant, the second simulation produced one cluster, and the third simulation produced two clusters. The backbone RMSD for each simulation and the central structures from the cluster analysis are shown in figures 8-10 with the unrestrained part shown in blue. The structure from simulation two (figure 9b) was the final structure accepted for further analysis as it represented a large number of structures. Also, upon visual inspection it was the most similar to the structure in cluster one from simulation three and the structure in cluster two from simulation one.

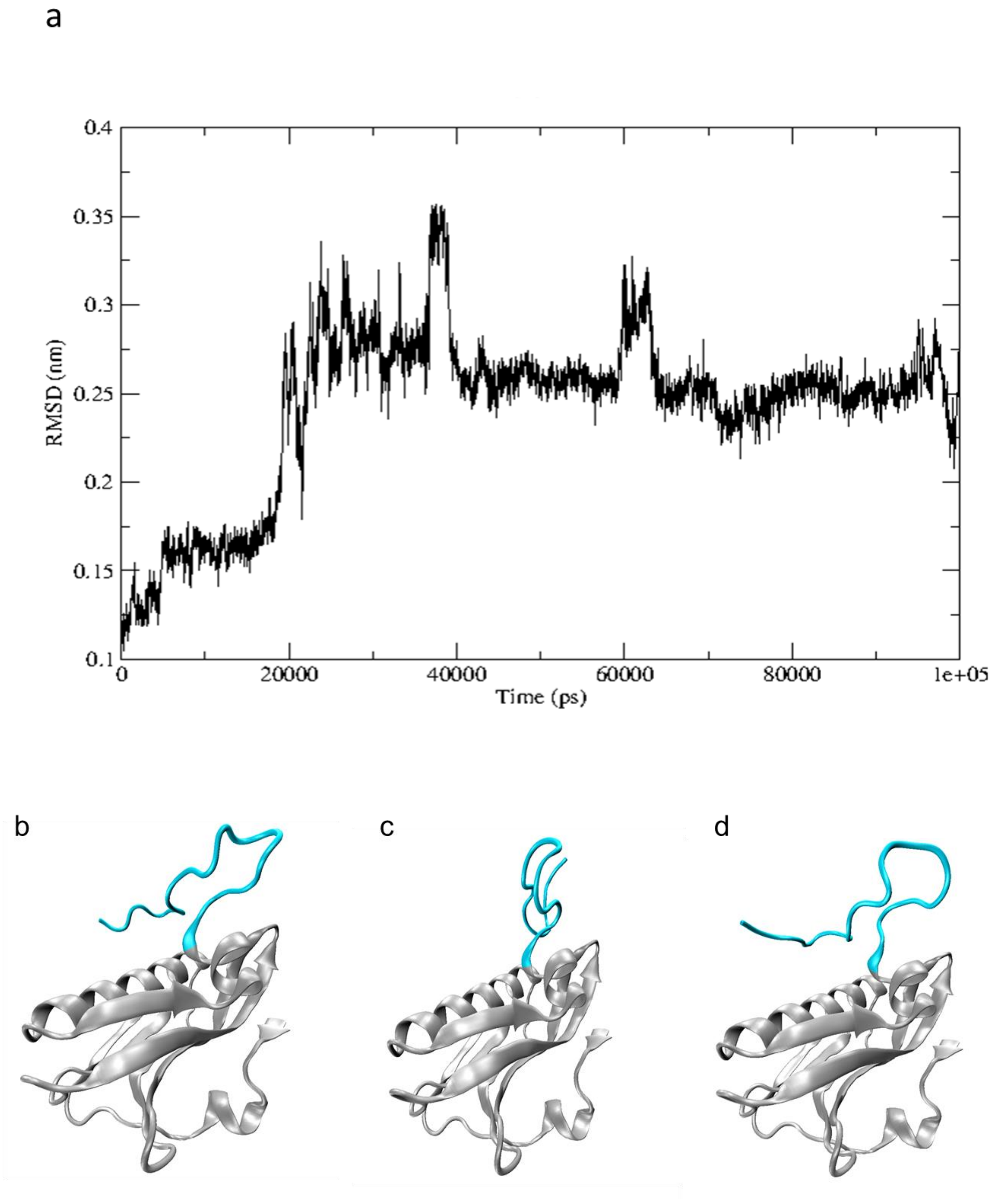


Figure 8. (a) RMSD plot of protein backbone coordinates for simulation one over 100 ns. For clustering analysis the trajectory was analysed from 20,000 ps. The RMSD cut-off was set at 0.2 nm. (b) C-terminus conformation (blue) of the central structure from cluster one representing 96% of the analysed trajectory. (c) C-terminus conformation (blue) of the central structure from cluster two representing 3.7% of the analysed trajectory. (d) C-terminus conformation of the central structure from cluster three representing 0.32% of the analysed trajectory.

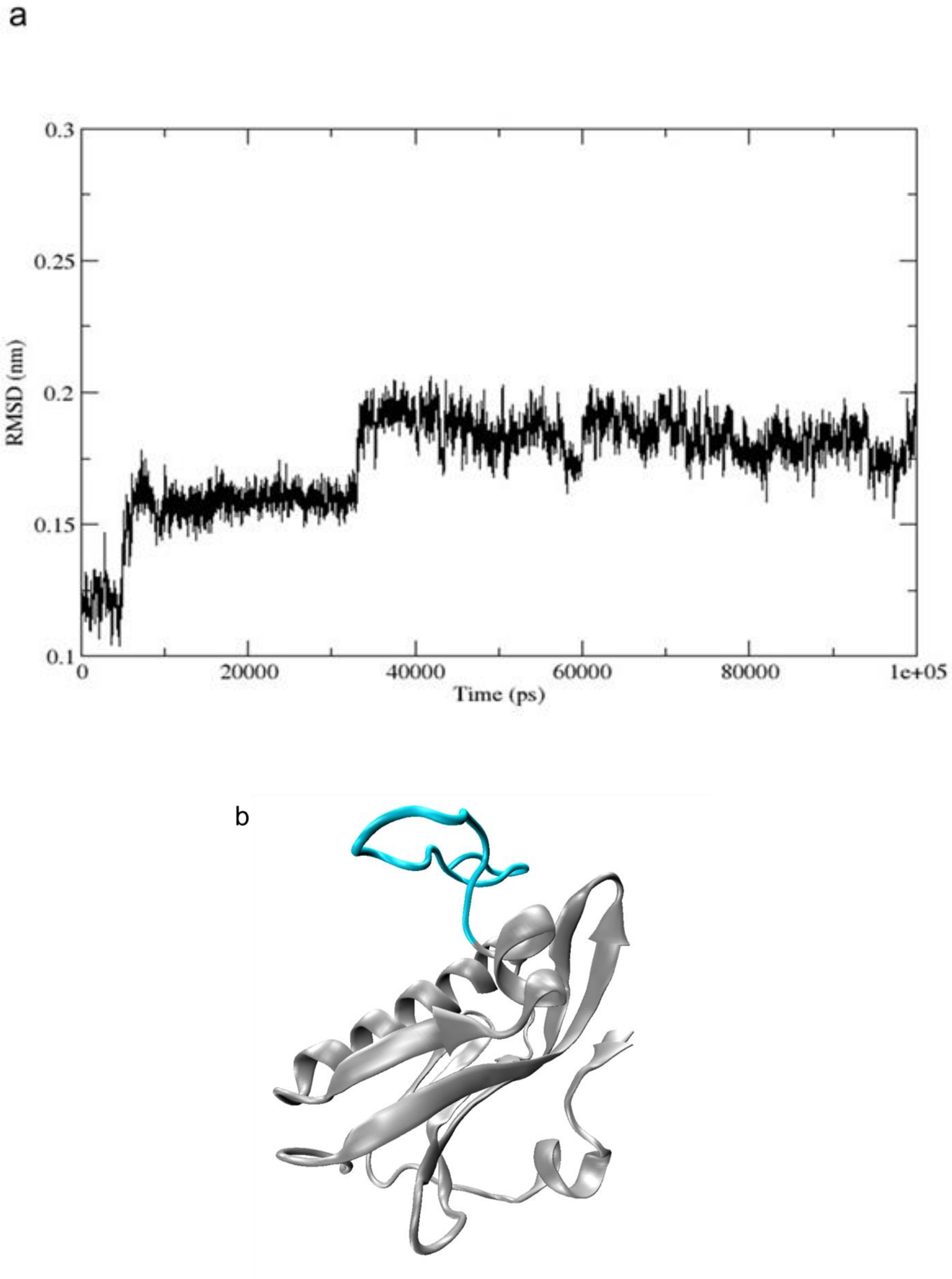


Figure 9. (a) RMSD plot of protein backbone coordinates for simulation two over 100 ns. For clustering analysis the trajectory was analysed from 35,000 ps. The RMSD cut-off was set at 0.15 nm. (b) C-terminus conformation (blue) of the central structure from cluster one representing 100% of the analysed trajectory.

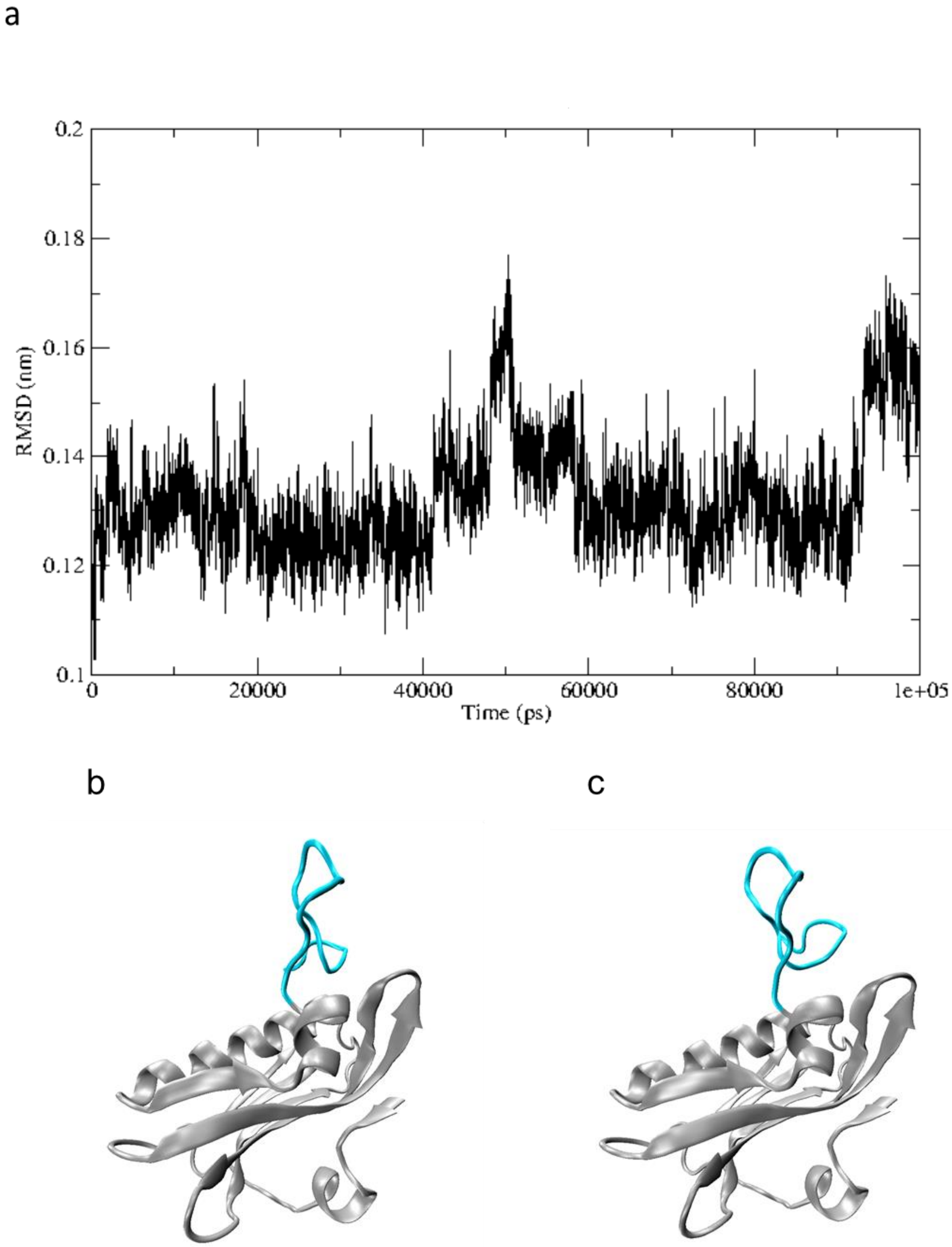


Figure 10. (a) RMSD plot of protein backbone coordinates for simulation three over 100 ns. For clustering analysis the trajectory was analysed from 40,000 ps. The RMSD cut-off was set at 0.1 nm. (b) C-terminus conformation (blue) of the central structure from cluster one representing 99.9% of the analysed trajectory. (c) C-terminus conformation (blue) of the central structure from cluster two representing 0.07% of the analysed trajectory.

3.4.3 Computational solvent mapping

Sixteen potential binding hot spots were identified through computational solvent mapping using the FTMap server. FTMap assigned the hot spots in clusters based on the number of different molecular probes binding to the site and ranked accordingly, with the highest ranked site, 'site one', binding with the largest number of different probes. The top ten binding sites were found to be located in different regions of the effector domain. Only sites fourteen and fifteen were located on the RNA binding domain. Several sites were located close together such as one and six, two and three, and seven and eleven. All hot spots were close to the surface of the protein. The level of residue conservation together with the top ten binding site locations is shown in figure 11. A distance of 4.0 Å was chosen to indicate the amino acid residues surrounding the top five hot spots (table 5). It was decided to consider binding sites one and six as a larger consensus site and subject it to virtual screening as this region is highly conserved and also composed of residues known to interact with host cell proteins.

Table 5. NS1 amino acid residues within 4.0 Å of the five highest ranked binding sites predicted by FTMap.

Site	Number of probes bound	Residues within 4.0 Å of binding site
1	13	Trp97, Phe98, Leu100, Asp115, Ile118, Val152
2	9	Leu139, Ile140, Pro157, Pro159
3	10	Met93, Ser94, Ile140, Leu141
4	10	Arg135, Arg195, Asn200, Pro210
5	9	Ile155, Gly174, Ile177, Gly178, Glu181

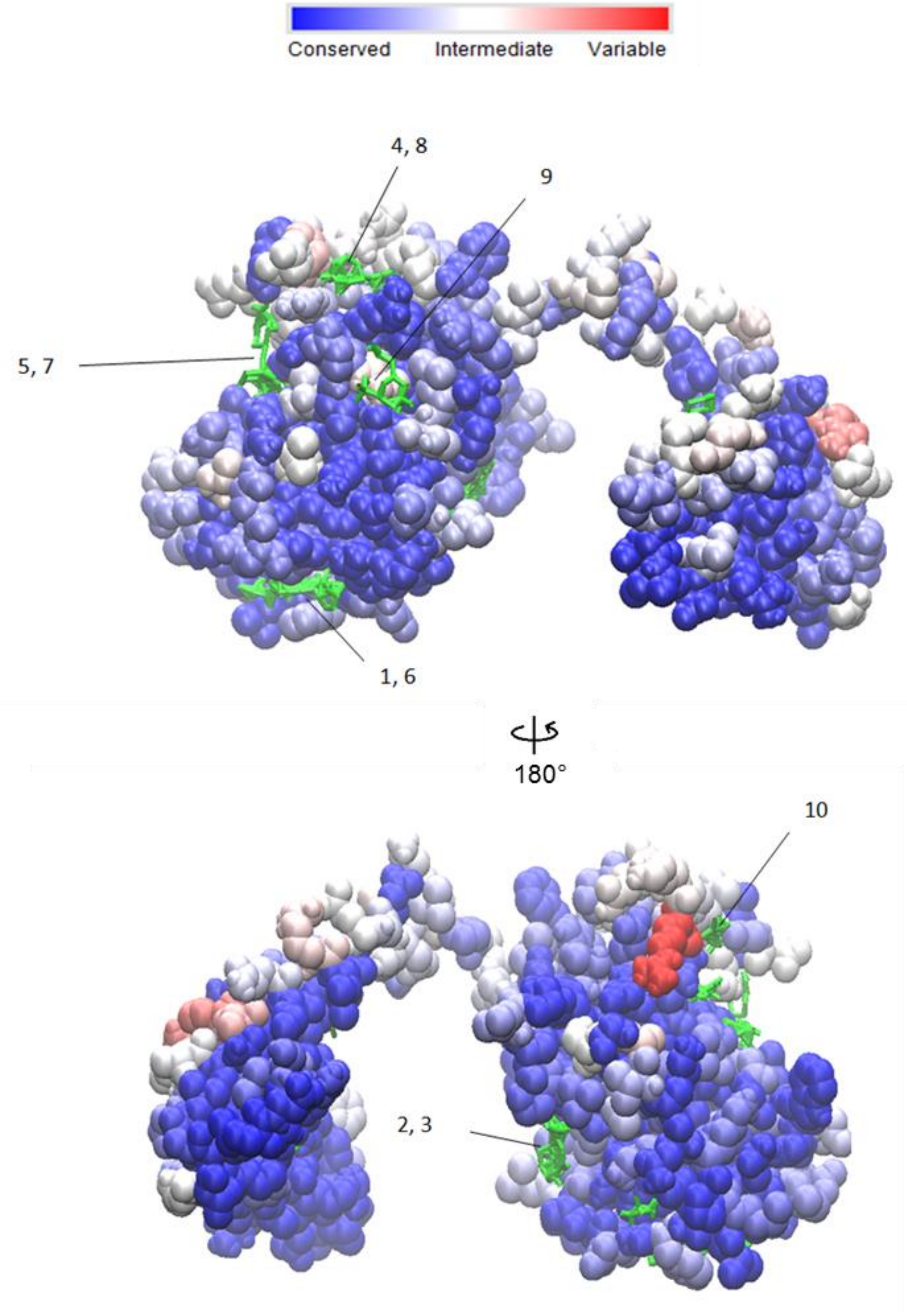


Figure 11. Ligand binding sites identified and ranked by the FTMap solvent mapping algorithm (green) and degree of conservation shown together on the NS1 structure. Blue regions indicate high conservation and red regions indicate low conservation (see colour scale).

3.4.4 Virtual screening

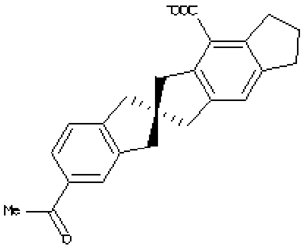
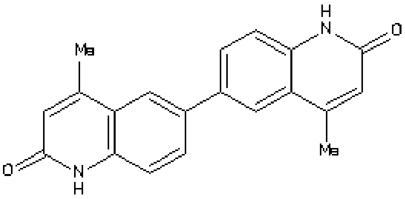
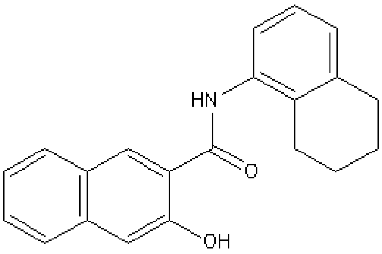
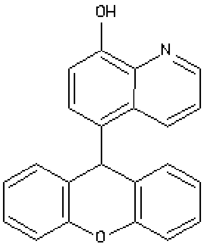
The results of the binding affinities for ~52,000 molecules docked against the NS1 target site using AutoDock 4 ranged from –10.9 kcal/mol to +2.7 kcal/mol, and -10.3 kcal/mol to -1.7 kcal/mol using AutoDock Vina. For the decoy site screening, the binding affinities using AutoDock 4 ranged from –7.5 kcal/mol to +357.3 kcal/mol, and from -7.1 kcal/mol to +28.8 kcal/mol using AutoDock Vina. The binding site rank lists were adjusted based on the top 15% of molecules in the decoy site rank list and then combined using the script in appendix 9.10.

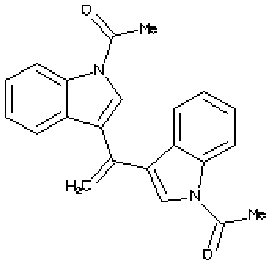
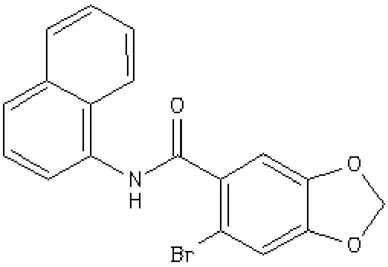
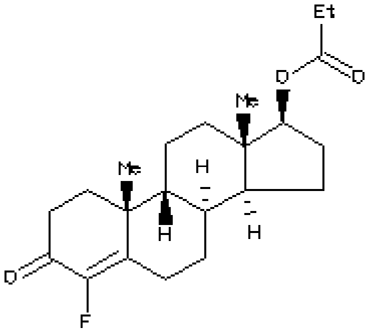
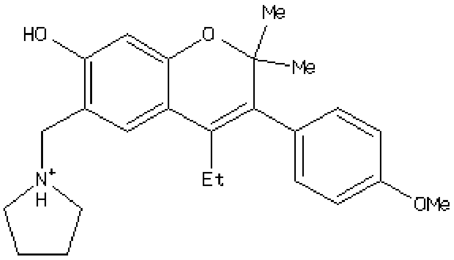
Fifteen molecules from the combined rank list were selected based on binding affinity and Lipinski's rule of five, which state that drug like compounds should ideally display the following properties (Lipinski, Lombardo, Dominy, & Feeney, 2001):

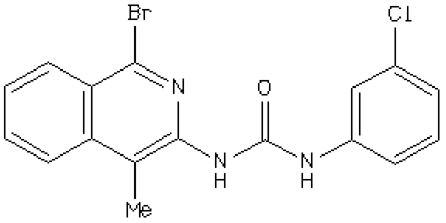
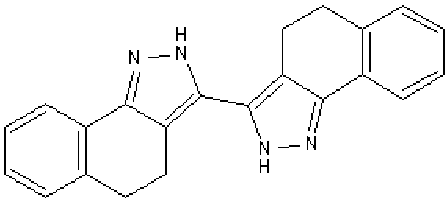
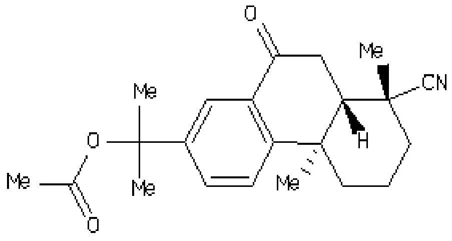
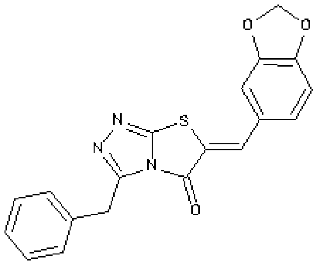
- Molecular mass less than 500 Daltons
- High lipophilicity (expressed as LogP less than five)
- Less than five hydrogen bond donors
- Less than ten hydrogen bond acceptors

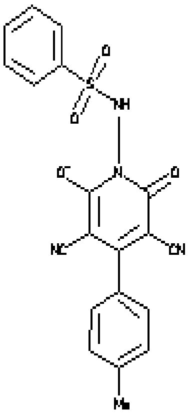
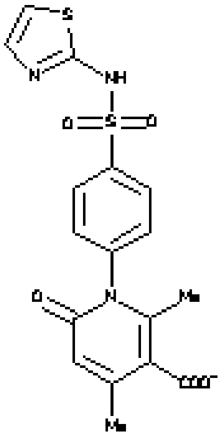
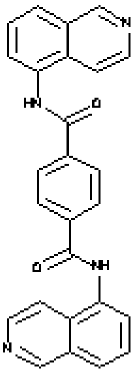
The chemical structures and properties of the top fifteen compounds selected and ordered for experimental testing are presented in table 6. The compounds were re-named alphabetically as indicated by the letters in brackets preceding the ZINC code. Common amino acid residues of the NS1 protein involved in binding these molecules identified using the software LigPlot+ (Laskowski & Swindells, 2011) include Phe98, Leu100, Asp115, Ala150 and Ile151, which are all well conserved. Docking models between four of the selected top molecules are shown in figure 12 and molecular interactions with compound I are shown in figure 13.

Table 6. Chemical structures of the top 15 compounds selected from the virtual screening based on binding affinity (ΔG), number of Hydrogen bond donors (H Don) and acceptors (H Acc), molecular weight and partition coefficient (xLogP) obtained from the ZINC database.

Compound (ZINC ID)	**ΔG (kcal/mol)**	**H Don**	**H Acc**	**Mol Wt (g/mol)**	**xLogP**
 (A) ZINC01646194	-8.50	0	3	345	3.75
 (B) ZINC16998207	-8.40	2	4	316	3.50
 (C) ZINC01669818	-8.10	2	3	317	5.60
 (D) ZINC04769085	-8.0	1	3	325	4.98

Compound (ZINC ID)	**ΔG (kcal/mol)**	**H Don**	**H Acc**	**Mol Wt (g/mol)**	**xLogP**
 (E) ZINC00344361	-7.90	0	4	342	4.83
 (F) ZINC01566093	-7.70	1	4	370	4.60
 (G) ZINC17002748	-8.90	0	3	362	4.34
 (H) ZINC01760393	-8.80	2	4	395	5.22

Compound (ZINC ID)	ΔG (kcal/mol)	H Don	H Acc	Mol Wt (g/mol)	xLogP
 (I) ZINC13281542	-8.60	2	4	391	5.56
 (L) ZINC05479272	-8.50	2	4	338	3.80
 (K) ZINC06761478	-8.49	0	4	353	4.64
 (O) ZINC01625245	-8.40	0	6	363	3.63

Compound (ZINC ID)	ΔG (kcal/mol)	H Don	H Acc	Mol Wt (g/mol)	xLogP
 (J) ZINC01650972	-8.50	1	8	405	2.78
 (N) ZINC01580219	-8.40	1	8	404	1.86
 (M) ZINC01703250	-8.44	2	6	418	3.85

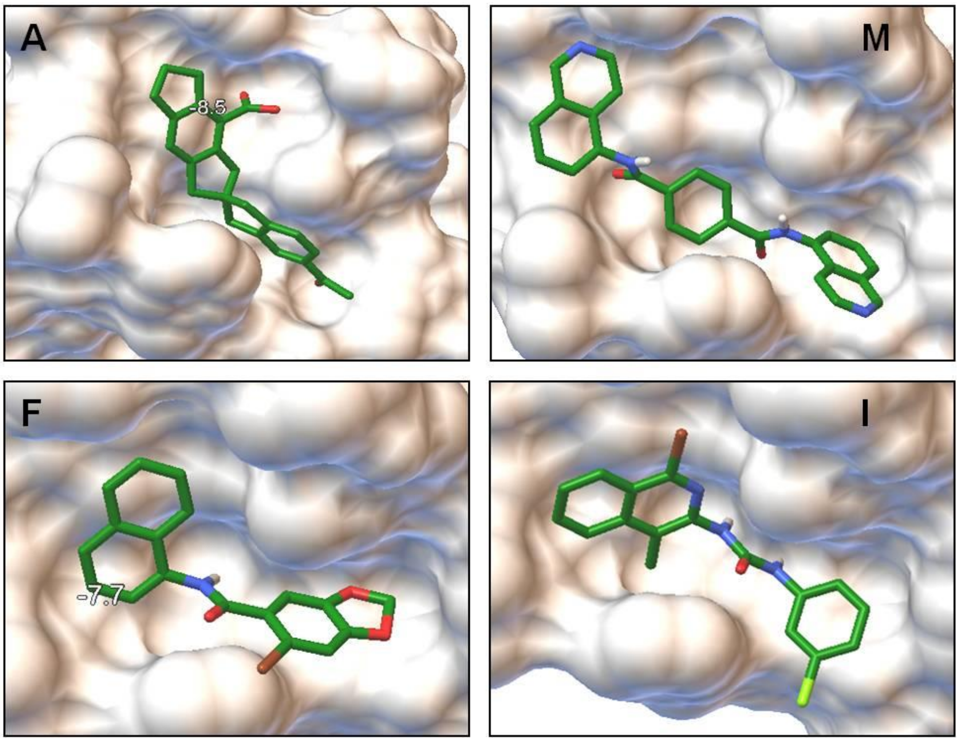
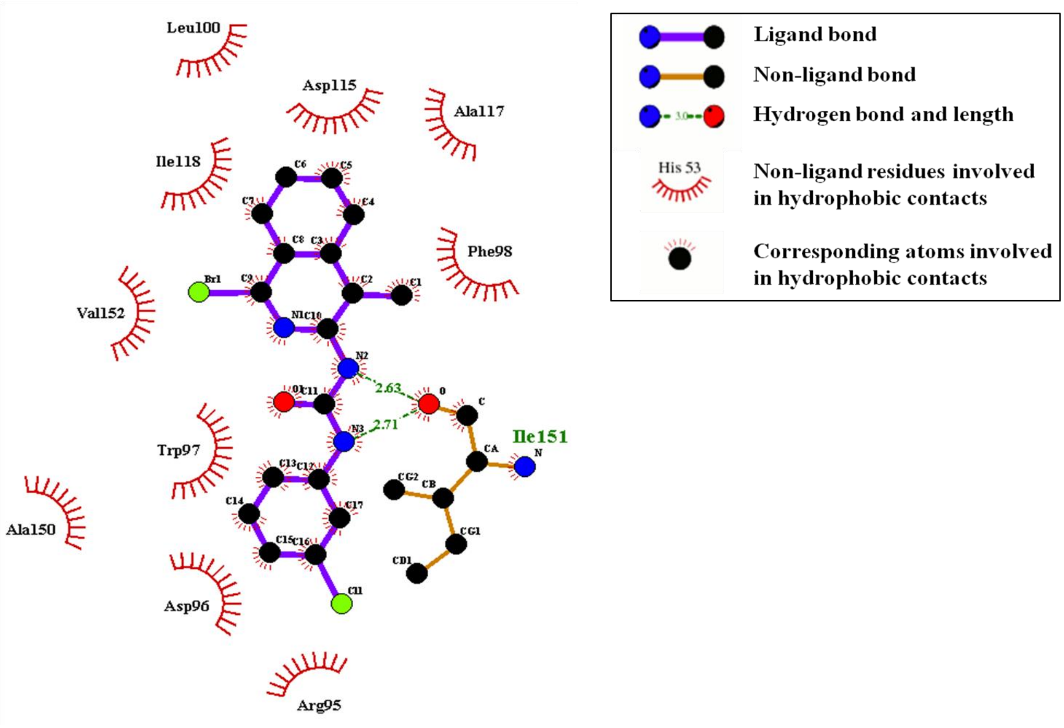


Figure 12. Models of four predicted top hit compounds (A, F, M, I) docked to the target site of the NS1 protein.



NS1_ZINC13281542

Figure 13. Molecular interactions between compound I and amino acid residues of the NS1 target site after docking with AutoDock 4.

3.4.5 Plaque assay

The virus dilution resulting in a countable number or visible plaques in the MDCK-SIAT cell line was x10^{-4}, and the mean number of plaques at this dilution was 81. The number of plaque forming units per ml (PFU/ml) at this dilution was calculated as: 81 x 10^4 x 10 = 8.1x10^6 PFU/ml.

3.4.6 Plaque reduction assay

Compounds I, L and M failed to dissolve in water, 100% DMSO, 100% ethyl acetate and 100% ethanol, and therefore were not tested *in vitro*. The average PFU/ml in cells treated with 0.1% DMSO and infected with a x10^{-4} virus dilution was calculated as: ~112 x 10^4 x 10 = ~1.12x10^7 PFU/ml by standard plaque assay, although this varied slightly between each assay.

The staining of the cell control (0.1% DMSO treated) showed that the cells had lasted under the experimental conditions over the five day period. The plaque morphology at each drug concentration was compared by visual observation to the virus control at day three post infection. Results for Oseltamivir carboxylate (positive control, figure 14) showed a distinct reduction in plaque size with increased drug concentration in each run, confirming that the assay was working, with the lowest concentration to reduce the plaque size by 50% being 0.1 µM. Out of the twelve remaining test compounds, most showed no success at changing plaque morphology and inhibiting virus replication. Several compounds (A, C, D, E, F and G) were toxic to the cell monolayer at 100 µM concentration.

However, compound D (figure 15) and K (figure 16) did show a reduction in plaque size at different concentrations. (Refer to figure 5 for plate layout). Compound D showed a weak inhibitory effect at 10 µm, 1.0 µM, 0.1 µM and 0.01 µM. Compound K showed inhibition between 100 µM and 10 µM, therefore was also tested with a two-fold dilution from 100 µM to 6.25 µM to narrow down the most effective inhibitory concentration; although it was observed that the change in plaque size was very similar at each intermediate concentration.

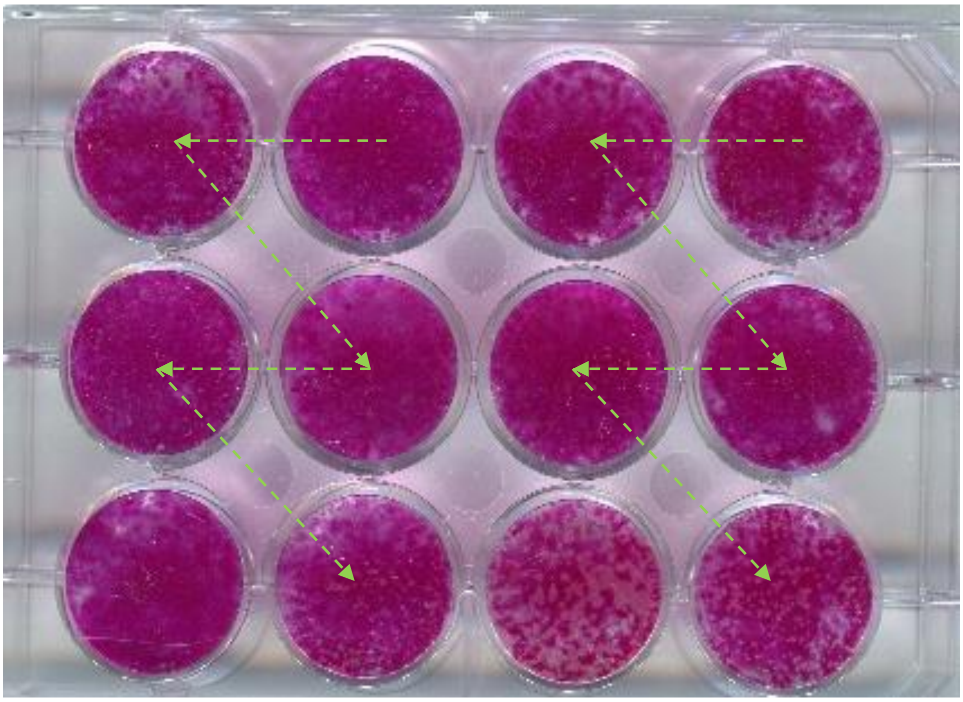


Figure 14. Plaque reduction analysis of Oseltamivir carboxylate on MDCK-SIAT cells infected with influenza A/Eng/529/2013 under agar overlay. A ten-fold drug dilution series was overlaid in duplicate from 100 µM to 0.01 µM (following the green arrow) and the bottom left well of each duplicate shows the controls (without drug).

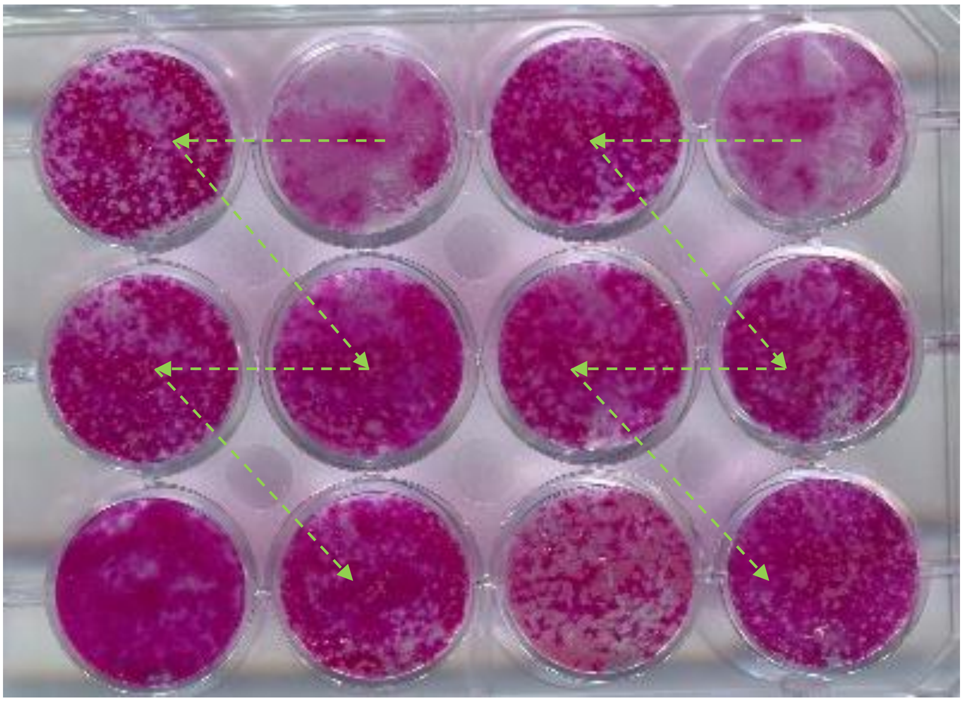


Figure 15. Plaque reduction analysis of compound D on MDCK-SIAT cells infected with influenza A/Eng/529/2013 under agar overlay. A ten-fold drug dilution series was overlaid in duplicate from 100 µM to 0.01 µM (following the green arrow) and the bottom left well of each duplicate shows the controls (without drug).

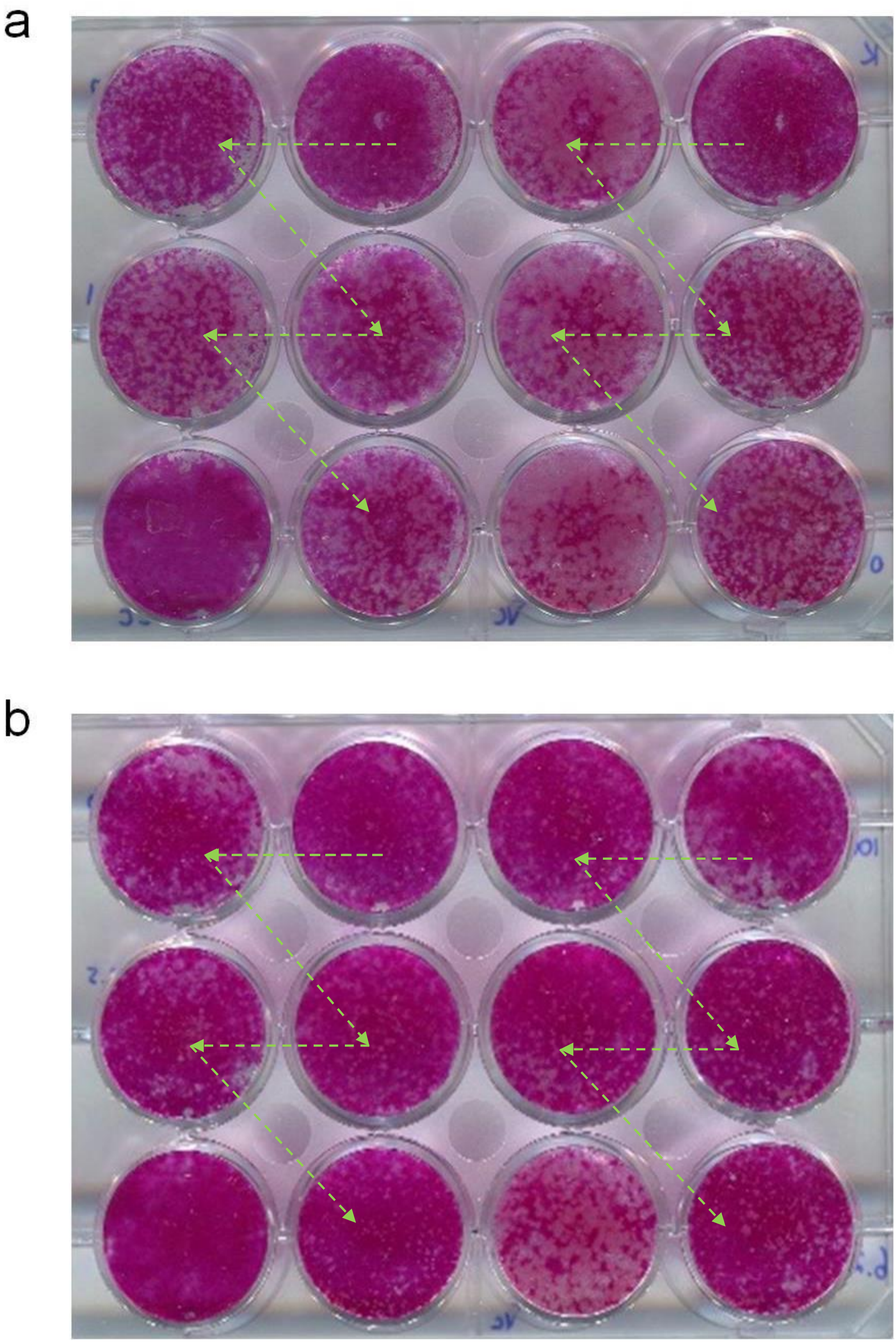
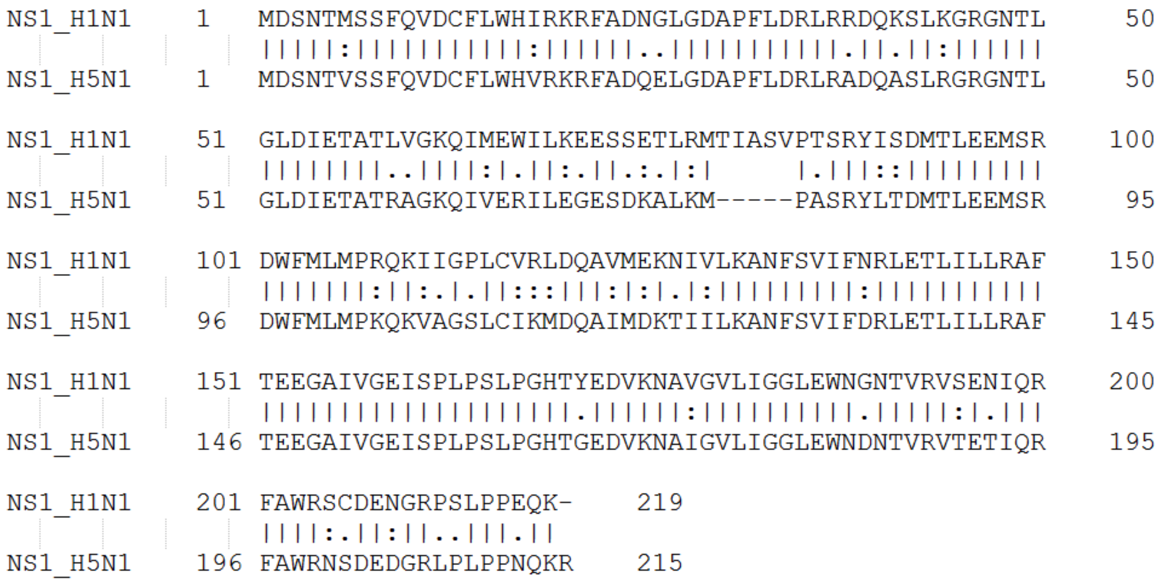


Figure 16. Plaque reduction analysis of compound K on MDCK-SIAT cells infected with influenza A/Eng/529/2013 under agar overlay. (a) Ten-fold drug dilution overlaid in duplicate from 100 µM to 0.01 µM (b) Two-fold drug dilution overlaid in duplicate from 100 µM to 6.25 µM (following the green arrow) and the bottom left well of each duplicate shows the controls (without drug).

3.4.7 H5N1 and H1N1 sequence alignment

The pairwise sequence alignment between the H5N1 sequence used for molecular modelling and the H1N1 sequence used for antiviral experiments is shown in figure 17. The two sequences (both isolated from humans) are not identical and differ by 49 amino acids. The overall percent identity between the two sequences is 78% and the overall percent similarity is 88% calculated using the EBLOSUM62 matrix. Most notably, the H1N1 sequence does not have the five amino acid deletion (T, I, A, S, V) at position 80 in the linker region, unlike the H5N1 sequence. Several amino acids which were predicted to interact with the top hit molecules (e.g Leu105, Asp120, Ala155) are identical between the two sequences.



```
NS1_H1N1     1   MDSNTMSSFQVDCFLWHIRKRFADNGLGDAPFLDRLRRDQKSLKGRGNTL     50
                 |||||:||||||||||||:||||||..|||||||||||.||.||:||||||
NS1_H5N1     1   MDSNTVSSFQVDCFLWHVRKRFADQELGDAPFLDRLRADQASLRGRGNTL     50

NS1_H1N1     51  GLDIETATLVGKQIMEWILKEESSETLRMTIASVPTSRYISDMTLEEMSR    100
                 ||||||||..||||:|.||:.||.:.|:|     |.|||::|||||||||
NS1_H5N1     51  GLDIETATRAGKQIVERILEGESDKALKM-----PASRYLTDMTLEEMSR     95

NS1_H1N1     101 DWFMLMPRQKIIGPLCVRLDQAVMEKNIVLKANFSVIFNRLETLILLRAF    150
                 |||||||:||:.|.||:::|||:|:|.|:|||||||||:|||||||||||
NS1_H5N1     96  DWFMLMPKQKVAGSLCIKMDQAIMDKTIILKANFSVIFDRLETLILLRAF    145

NS1_H1N1     151 TEEGAIVGEISPLPSLPGHTYEDVKNAVGVLIGGLEWNGNTVRVSENIQR    200
                 ||||||||||||||||||||.||||||:||||||||||.|||||:|.|||
NS1_H5N1     146 TEEGAIVGEISPLPSLPGHTGEDVKNAIGVLIGGLEWNDNTVRVTETIQR    195

NS1_H1N1     201 FAWRSCDENGRPSLPPEQK-    219
                 ||||:.||:||..|||.||
NS1_H5N1     196 FAWRNSDEDGRLPLPPNQKR    215
```

Figure 17. NS1 human H5N1 (A/Vietnam/1203/2004) and human H1N1 (A/England/529/2013) pairwise sequence alignment output. A vertical line indicates positions which have a single fully conserved residue, a colon indicates strongly similar properties and a period indicates weakly similar properties. (Ala at position 38 and 41 of the H5N1 sequence is mismatched to Arg38 and Lys41 in the Uniprot sequence).

3.4.8 Drug Affinity Responsive Target Stability (DARTS)

The binding of compound D at a concentration of 10 µM to the NS1 protein was tested based on the ability of compound D to stabilise NS1 against limited proteolysis. The proteolysis results using three different pronase concentrations (diluted from a starting concentration of 1.25 mg/ml) are shown on the western blot in figure 18. The strong bands produced at 26 kDa confirm presence of the NS1 protein in the cell lysate. At the highest pronase dilution of 1:10000, there is more proteolysis in the sample with the addition of the drug compound D (lane 3) than with DMSO vehicle control (lane 7). The detection of bands shows that the DARTS method is unable to confirm that compound D binds to the NS1 protein.

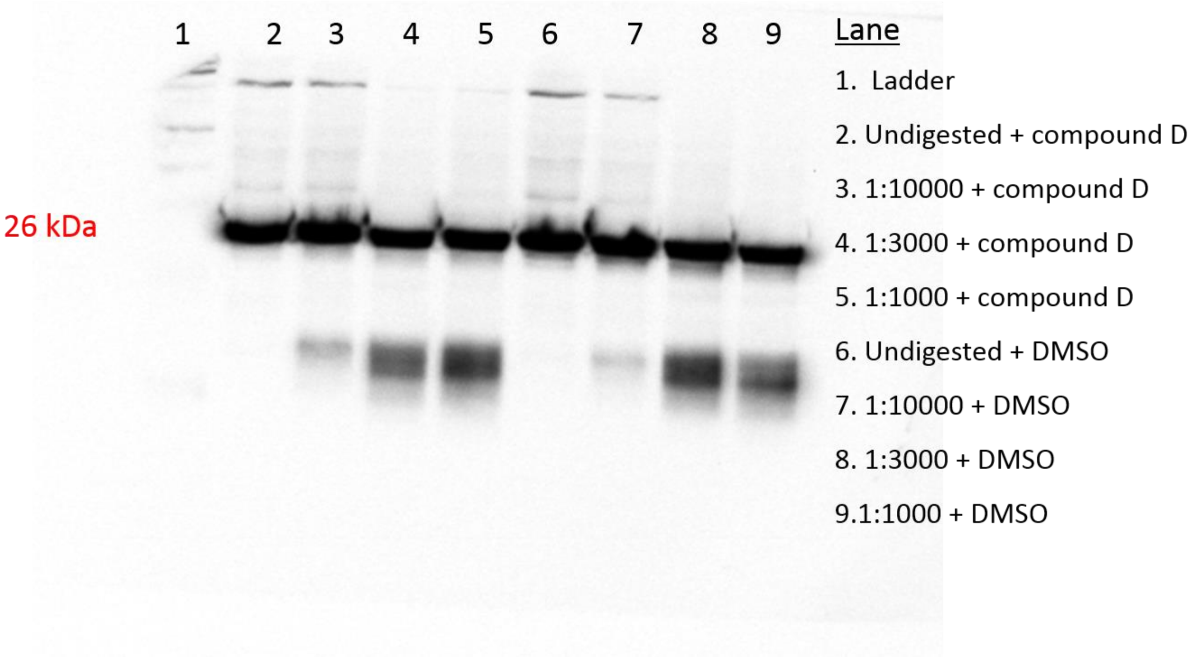


Figure 18. Western blot showing the amount of influenza A NS1 proteolysis using three different pronase concentrations after incubation with drug compound D and with DMSO as vehicle control.

3.4.9 Potential drug target identification - PharmMapper

As the DARTS method was unable to confirm binding of compound D to the NS1 protein, it was tested, if compound D is predicted to bind to any other drug target for which pharmacophore models are available. The top five pharmacophore models identified by the PharmMapper web server are shown in table 7. The top target identified (sorted by Z-score) was bacteriorhodopsin which interacts through hydrophobic interactions with compound D.

Table 7. PharmMapper results showing five top hit pharmacophore models predicted and their feature interactions.

Rank	Target Name	PDB ID	Z-score	No. hydrophobic	No. HB acc	No. HB donors
1	Bacteriorhodopsin	1P8U	3.24	7	0	0
2	Vanillyl-alcohol oxidase	1W1J	2.51	2	2	0
3	Heat shock protein HSP 90-alpha	1YC1	2.07	2	3	2
4	Thyroid hormone receptor beta	1N46	2.02	4	1	1
5	Thyroid hormone receptor beta	1NAX	2.01	4	1	1

3.5 DISCUSSION

3.5.1 NS1 conservation

The degree of NS1 protein conservation was determined from sequence data from all hosts and subtypes until 2014. The protein is highly conserved, as over 85% of residues had a conservation score of 0.800 or above based on the Valdar scores obtained from the multiple sequence alignment. The RNA binding domain (RBD) is a well conserved region, with helices one and two consisting of highly conserved residues; while positions 55, 56, 59 and 60 of helix three display intermediate or low conservation. Also, a short stretch of less conserved residues between positions 21-26 form the loop region connecting helix 1 and 2. Regions on the effector domain such as the nuclear export signal region (residues 133-145 (Tynell, Melén, & Julkunen, 2014)) were highly conserved with scores >0.800, as well as Trp182 with a score of 0.998 which is most likely related to its importance for ED homodimerisation (Ayllon, Russell, García-Sastre, & Hale, 2012).

The large variability in residue composition, and length of the C-terminal tail and linker region, which is frequently reported in literature (Hale et al., 2010; Tu et al., 2011) was observed from the sequence alignment. Regions of low conservation may indicate that residues at that position do not play a significant role in the proteins overall function. Alternatively, major differences between sequences may result in changes in protein structure or biological functions. An example of this is the 205-237 variable region, where sequences with extended C-terminal tails are able to bind the poly A binding protein II (PABPII) causing nuclear accumulation of cellular RNA (Chen et al., 1999), although this interaction is probably not essential for the overall viral life cycle, especially as many sequences display C-terminal truncations. Regarding the linker region, it is known that structural changes in domain orientation result from the variability of the residues present as well as determining NS1 subcellular localisation (Li, Noah, & Noah, 2011).

Compared to previous conservation analysis (Darapaneni et al., 2009), the results also show that despite the inclusion of many more protein sequences (~6000), the overall pattern of NS1 conservation is generally quite similar. Additionally, the different scoring method used in the current work to quantify conservation from a sequence alignment influences the classification into conserved or variable residues, which is seen with a small number of residues in this analysis.

3.5.2 Predicted binding hot spots

Sixteen binding hot spots were predicted using FTMap, many of which were in conserved areas of the NS1 protein structure, with the top ten being located on the effector domain (ED). Compared to the RNA binding domain, this domain is known to interact with many different cellular proteins, therefore it is expected to have more binding sites. The clustered hot spots are all surface accessible, which facilitates ligand binding. On the effector domain, binding site four and eight were shown to be spatially close together (within 5.0 Å), as well as sites five, seven and eleven.

Hot spots one, five, six and twelve consist of residues 144-186 that were reported to interact with the cellular host protein CPSF30 (Twu et al., 2006). A later structural study of the ED based on the crystal structure from a H3N2 subtype (A/Udorn/72) defined the F2F3 domains of the CPSF30 binding pocket to be made up of residues Lys110, Ile117, Ile119, Glu121, Val180, Gly183, Gly184 and the highly conserved Trp187 (Das et al., 2008). Furthermore, the highly conserved Ile64 located on the RBD has recently been identified to interact with CPSF30 and its mutation to Thr64 (which occurs rarely) decreases this interaction (DeDiego, Nogales, Lambert-Emo, Martinez-Sobrido, & Topham, 2016). The biological consequence of a ligand binding to these residues may prevent the inhibitory activity on host pre-mRNA processing. Although some of these residues including Trp187 (position 182 in the H5N1 A/Vietnam/2004 sequence) which prominently protrudes from the protein and is required for ED dimer formation (Ayllon et al., 2012; Carrillo et al., 2014) were not close to any predicted binding sites.

Also, the highly conserved glutamic acid residues at positions 91 and 92 which form the TRIM25 binding domain required for suppressing interleukin-1β secretion were not found to be close to any predicted binding sites (Gack et al., 2009; Moriyama et al., 2016), whilst site two, three and four neighbour residues between positions 123-144 which may interact with the host protein hGBP1, which has a role in antagonising antiviral activity (Zhu et al., 2013). Site ten and eleven are located closest to the flexible and variable C-terminal region of the protein. Due to the high conservation and rank assigned by the FTMap server, site one and six were selected as the target site for virtual screening.

3.5.3 Top hit compounds and antiviral activity

The consensus virtual screening and docking predictions showed that a large number of compounds have high predicted binding affinities to the NS1 target site. However, many of these may be false positive predictions (Chen, 2015). The docking conformations based on the output coordinate files enabled critical residues which directly participate in molecular interactions to be identified. Some of the top compounds shown in table 6 were predicted to interact with Phe98, Leu100, Asp115, Ile118, Lys121, Glu148, Ala150 and Ile151, all of which are well conserved. These interactions may involve, hydrogen bonding, hydrophobic interactions and ionic interactions.

The plaque reduction assay, which is a broad phenotypic assay suitable for use when virus-drug target interactions are unknown, was performed to screen and determine the effect on influenza A replication after incubation with a compound at five different concentrations. Two out of twelve compounds tested (D and K) showed capability of reducing H1N1 virus replication at non-cytotoxic concentrations. This is a successful hit rate and is comparable with other experimental screening studies previously reported, such as the identification of six possible NS1 binding inhibitors from an initial screen of 446 small molecules after development of a fluorescence polarisation-based high throughput screening assay (Cho et al., 2012), and the discovery of four compounds out of an initial screen of ~2000 that showed influenza antiviral activity, which are thought to suppress NS1 function (Basu et al., 2009). The virtual screening approach in this work also presents the advantage of saving costs compared to using experimental screening methods.

The remaining ten compounds showed no significant visible effect, although it is possible that these compounds and compound K, may still bind to NS1, albeit with no overall antiviral mechanism of action. Additionally, it has been shown with the example of PB2 inhibitors that the antiviral activity of the inhibitor compound may depend on the bioassay used (Stevaert & Naesens, 2016).

The binding of compound D to the NS1 protein was investigated with the DARTS method, that is based on the assumption that binding of a ligand protects the protein from proteolysis (Lomenick et al., 2011). The results indicate that the experiment was conducted successfully, as the primary antibody was able to detect the NS1 protein and proteolysis had occurred at different pronase concentrations as shown on the Western blot (figure 18). However, the DARTS method did not show that the NS1 protein was protected from proteolysis by compound D, even under limited proteolysis conditions. The Western blot indicated that a similar level or even more proteolysis had occurred in the sample incubated with the drug, compared to the sample without. This could be because compound D does not bind to NS1, the concentration of compound D may have been too low, or that binding of compound D does not increase the stability of NS1 to protect it from proteolysis, hence the detection of protein bands below 26 kDa.

A computational drug target prediction with PharmMapper showed that compound D does not bind major cellular target proteins, but it has the potential to bind to hydrophobic parts of transmembrane proteins such as bacteriorhodopsin. It may even interact with multiple targets. This may be an alternative explanation of the antiviral activity and it is possible that compound D probably binds to and inhibits a different influenza A protein, such as the transmembrane M2 protein, or haemagglutinin that is essential for the membrane fusion process. Overall, the results showed that through experimental testing of twelve compounds following virtual screening of ~50k compounds, an active inhibitor could be identified, although the predicted mechanism of action could not be confirmed. Additionally, the choice of compounds for testing is challenging as compounds with ideal drug like properties and strong binding affinity may not be compatible with cellular uptake.

4 THE NUCLEAR EXPORT PROTEIN (NEP)

4.1 INTRODUCTION

The nuclear export protein (formerly known as the non-structural protein 2 or NS2) is encoded by RNA segment eight and is 121 amino acid residues long. In the virion it is present bound to the M1 matrix protein through interaction with the prominently surface exposed residue Trp78. This interaction facilitates the main function of the NEP which is to export newly synthesized viral ribonucleoproteins (vRNPs) from the nucleus to the cytoplasm of infected cells to allow translation of mRNA into the structural proteins, as well as packaging the genome into progeny virions (Akarsu et al., 2003). This is an essential process during the replication cycle and is mediated by binding with the cellular export molecule chromosome region maintenance protein 1 (CRM1) and its cofactor ranGTP (Neumann, Hughes, & Kawaoka, 2000). A model of vRNP export mediated by the NEP is shown in figure 19.

Many other key functions of the NEP have also been found, such as regulating accumulation of viral RNA during transcription and translation (Robb, Smith, Vreede, & Fodor, 2009), interacting with cellular molecules such as ATPase to assist with viral budding (Paterson & Fodor, 2012) and several nucleoporins to enable nucleocytoplasmic transport (Chen, Huang, & Chen, 2010; O’Neill, Talon, & Palese, 1998). The NEP has been found to be phosphorylated in infected cells (Richardson & Akkina, 1991), and the phosphoacceptor site(s) may be required for regulating nuclear export of vRNPs and/or polymerase activity (Reuther, Giese, Gotz, Riegger, & Schwemmle, 2014).

Previous work reporting structural information for the NEP have suggested that the protein consisting of four helices adapts a compact, yet flexible and highly mobile conformation, in particular the N-terminal fragment consisting of residues 1-53 (Darapaneni et al., 2009; Lommer & Luo, 2002). To date, there is only one experimentally determined structure available in the PDB (ID: 1PD3) for the NEP at a resolution of 2.6 Å (Akarsu et al., 2003). Furthermore, the N-terminal residues 1-62 did not crystallise, which strongly indicates structural disorder (Le Gall, Romero, Cortese, Uversky, & Dunker, 2007). The N-terminus is recognised by the nuclear export machinery during replication and contains a leucine rich nuclear export signal

(NES) formed by residues 11-23 (O’Neill et al., 1998). Ser17 and Leu21 in this region were found to be highly conserved upon large scale sequence analysis and are therefore proposed to be key residues of the NES (Darapaneni et al., 2009). A second NES in a leucine rich region between residues 22 and 45 has also been identified and is proposed to be involved in the export of vRNPs (Huang et al., 2013).

In contrast, the C-terminal region for which atom coordinates are available (Akarsu et al., 2003), is highly structured in the form of two co-linear α-helices and consists of several hydrophobic residues (Paterson & Fodor, 2012). The polymerase enhancing function of NEP has been shown to reside in the C-terminus (Reuther, Giese, Götz, et al., 2014). Unlike many of the other influenza A proteins, there have been no inhibitors identified which specifically target the NEP.

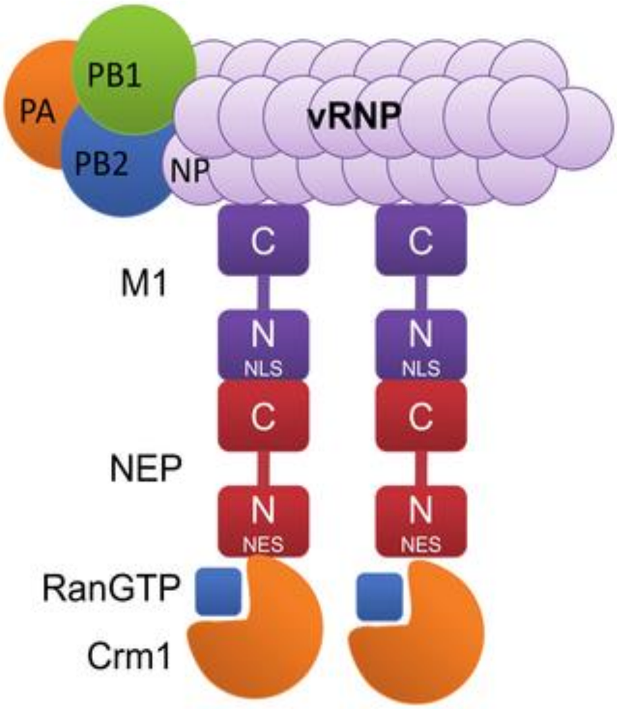


Figure 19. Model for NEP-mediated nuclear export of influenza viral ribonucleoproteins. The nuclear export signal of the NEP N-terminal domain is recognised by the cellular Crm1-RanGTP complex, whilst the C-terminal domain is bound to the viral M1 protein to facilitate translocation of vRNP’s towards the cytoplasm. Figure reproduced from Paterson & Fodor (2012).

4.2 METHODS

4.2.1 NEP sequence analysis

The sequence analysis was performed according to general methods section 2.2. Briefly, full length NEP sequences were chosen from all hosts, regions and subtypes until September 2015. Identical sequences and sequences containing non-standard residues were removed. The CD-HIT web server (Huang et al., 2010) was used to remove redundancy of the sequences at a similarity threshold of 98%. For calculation of amino acid conservation see general methods section 2.3. The minimum score used for re-scaling was 0.632.

4.2.2 Protein modelling

The crystal structure of the NEP M1 binding domain (PDB ID 1PD3) from a H1N1 sequence (A/Puerto Rico/8/1934) contained missing regions 1-62 at the N-terminus and 117-121 at the C-terminus. Therefore the I-TASSER server (Zhang, 2008) was used to generate a full model of this sequence. The I-TASSER model and chain A of the experimental structure 1PD3 were aligned by structural superposition in the protein structure viewer PyMol using the 'align' command and the coordinates of the alignment were saved as a PDB file. This PDB file was then modified to replace the structural information of residues 64-115 of the model with the experimentally known coordinates of 1PD3 using a text editor. Atom coordinates for residues 86 and 104 from the model were not replaced.

4.2.3 Molecular dynamics simulations

Coordinates of the full length NEP model were used for MD simulations using Gromacs 2016. The starting structure was converted to a Gromacs file and molecular topology file with the program pdb2gmx. The AMBER99SB – ILDN force field was selected as it demonstrated high accuracy in a systematic evaluation of eleven MD force fields (Beauchamp et al., 2012) along with the TIP3P water model. The protein was solvated in explicit water in a cubic box with periodic boundary conditions, and the charge of the chemical system was neutralised with five sodium ions and additional NaCl at 100 mM concentration. The following commands were used:

```
$ gmx_mpi pdb2gmx -f myNEP_2.pdb -o NEP.gro -p NEP.top -ter –ignh

$ gmx_mpi editconf -f NEP.gro -d 0.75 -o box.gro -bt cubic

$ gmx_mpi solvate -cp box.gro -cs spc216.gro -p NEP.top -o solvated.gro

$ gmx_mpi grompp -f ions.mdp -p NEP.top -c solvated.gro -o ions.tpr

$ gmx_mpi genion -s ions.tpr -neutral -conc 0.1 -p NEP.top -o ions.gro
```

Energy minimisation with 500 steps of the steepest descent algorithm followed by 100 ps MD simulation with position restraints on the protein non-hydrogen atoms were performed to equilibrate the system. The temperature of the system was set to remain constant at 300 K using the velocity rescaling thermostat with a coupling constant of 0.5 ps and pressure coupling was set at 1.0 atm with the Berendsen barostat. The Particle Mesh Ewald (PME) algorithm for calculating electrostatic interactions was used with the real space cut-off set at 1.0 nm with the Verlet cut-off scheme. The following commands were used:

```
$ gmx_mpi grompp -f em.mdp -p NEP.top -c ions.gro -o em1.tpr

$ gmx_mpi mdrun -ntomp 4 -s em1.tpr -c nepAfterEM.gro

$ gmx_mpi grompp -f posres.mdp -p NEP.top -c nepAfterEM.gro -o posresEq.tpr

$ gmx_mpi mdrun -ntomp 4 -s posresEq.tpr -c nepAfterPosResEqMD.gro -v >&
posres.log
```

For the production run, three replicate partially position restrained simulations over 100 ns (50 million steps) were performed. The same settings for equilibration were used, except the Parinello-Rahman barostat was used for pressure coupling and position restraints were released on N-terminus residues 1-65 by modifying the forces in the posre.itp file as shown in table 8. The simulations were performed at a time step of 2 fs. Coordinates were written to the compressed output trajectory file every ten thousand steps. The following commands were used:

```
$ gmx_mpi grompp -f partfree_md.mdp -p NEP.top -c nepAfterPosResEqMD.gro -t
state.cpt -o Sim1.tpr
```

```
$ gmx_mpi mdrun -s Sim1.tpr -c afterMD.gro -v -stepout 2000
```

Table 8. Implementation of position restraint forces in the x, y, z axis on the non-hydrogen atoms of the NEP protein. A force of 1000 indicates full restraint on an atom and 0 indicates full motion.

Atoms	Residue	fx	fy	fz
1-971	1-62	0	0	0
972-1032	63-65	500	500	500
1033 onwards	66 -121	1000	1000	1000

Simulation trajectories were analysed by creating an index file specifying backbone atoms for residues 1-65 to calculate the root mean square deviation for the unrestrained part of the protein using the Gromacs function 'rms'. The implementation of position restraints was verified by analysing the root mean square fluctuation for each residue using the Gromacs function 'rmsf'. For cluster analysis the group selected for least squares fitting and RMSD calculation was residues 1-65, and the group for output was selected as protein.

4.2.4 Prediction of intrinsic disorder

A top ranking intrinsic disorder prediction server DISOPRED3 (http://bioinf.cs.ucl.ac.uk/psipred/?disopred=1) was used to predict any intrinsically disordered regions for the NEP structure. The server is based on pattern recognition of an amino acid sequence against disordered regions in the PDB (Jones & Cozzetto, 2015; Ward, McGuffin, Bryson, Buxton, & Jones, 2004). The full H1N1 sequence A/Puerto Rico/8/1934 used for modelling was entered as the input sequence.

4.2.5 Prediction of binding hot spots

Following cluster analysis of the three MD trajectories, the central structures from the largest simulation clusters were selected for solvent mapping to predict binding hot spots. See general methods section 2.5.

4.2.6 Virtual Screening

A binding hot spot predicted in the same location between helix two and four on each NEP structure following cluster analysis was selected as the target site for virtual screening. The grid box parameters for docking using AutoDock Vina and AutoDock 4 were set using AutoDock Tools as follows, with a grid spacing of 0.375 Å:

```
center_x = 11.756
center_y = 5.141
center_z = -0.693
size_x = 15.75
size_y = 16.5
size_z = 14.25
exhaustiveness = 12
```

The same NCI compound library used for screening against the NS1 protein was filtered using the software Open Babel version 2.3.1 (O'Boyle et al., 2011) to eliminate compounds with molecular weight over 500 g/mol and predicted logP over five using the command:

```
$ obabel output.mol2 --filter “MW<500 logP<5” –O filtered.mol2
```

The library was converted to SDF format using Open Babel to be passed through the Pan Assay Interference Compounds (PAINS) filter with the online FAF-Drugs3 (Free ADME-Tox Filtering Tool) program to identify and remove compounds that appear as frequent hitters in screening experiments (Baell & Holloway, 2010; Lagorce, Sperandio, Baell, Miteva, & Villoutreix, 2015). The filtered library was converted back to .mol2 format and compounds were split into individual ligand files in PDBQT format using the AutoDock screening preparation tool Raccoon. The DrugBank-approved compound library (version 4.0) containing 1738 drugs between pH 6 and 8 (Law et al., 2014) was downloaded from the ZINC database and also screened against the NEP target site in order to find any drugs which have been approved that may also target the NEP.

4.3 RESULTS

4.3.1 NEP amino acid conservation

3000 NEP sequences mainly from human, swine and avian hosts from all geographic regions were initially obtained from the NCBI Influenza Virus Resource (Bao et al., 2008). Applying a similarity threshold of 98% the redundancy was reduced with CD-HIT resulting in 889 sequences remaining. Clustal Omega was used to align the sequences and the alignment profile showed that sequences had a high level of conservation*.* The conservation scores for each amino acid based on Valdar's scoring method are presented in table 19 (appendix 9.2) and ranged from 0.632 at position 89 (lowest) to 1 (highest). Gly30, Leu38 and Arg66 were 100% conserved, whilst other highly conserved residues (>0.900) include Met16, Ser17, Tyr41, Leu69, Arg84 and Leu103. The most conserved region is from Thr90 to Glu110 which forms helix four at the C-terminus. The least conserved residues (<0.750) were found to be: Leu14, Glu22, Gly26, Ser57, Asn60, Glu63 and Ile89; several of these residues form the interhelical loops and turns between the helices. Frequent amino acid substitutions observed for positions displaying low conservation include E, M, R, Q, K, R, V, T and A at position 14, and K, T, V, I, N, L, A, S and T at position 89. For display purposes the scores were re-scaled and mapped onto the NEP structure shown in figure 20.

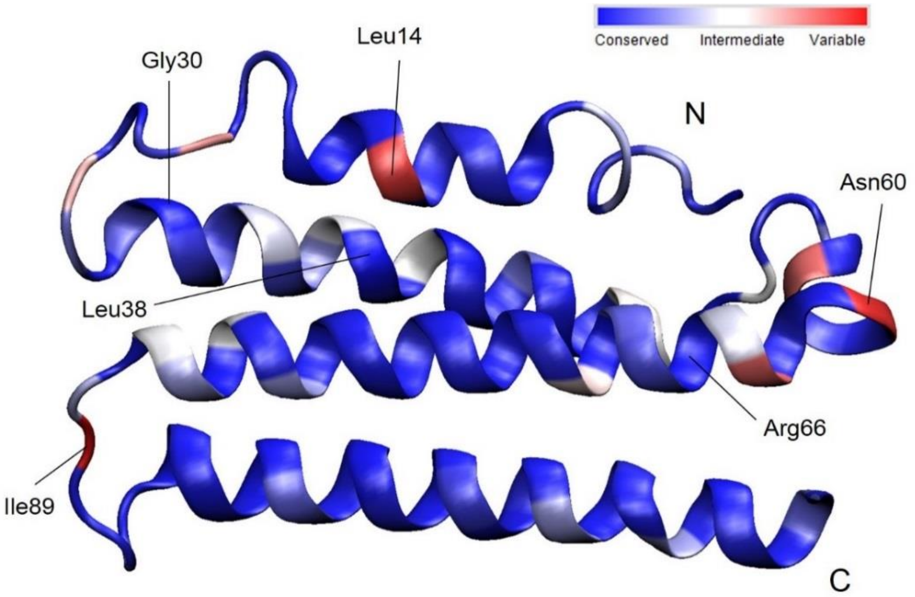


Figure 20. NEP conservation mapped onto the NEP structure. Blue regions indicate high conservation and red regions indicate low conservation.

4.3.2 Protein modelling and MD simulations

The I-TASSER server generated a full length NEP model based on the H1N1 input sequence. The C-terminal part of the experimental structure 1PD3 was used as the template and the remaining structure was modelled *ab-initio* as no suitable template was available. The RMSD between the C-terminal region of the model and 1PD3 was 1.98 Å, indicating a small difference between the two structures. The experimentally known coordinates of 1PD3 chain A (C-terminal residues 64-115) were combined with the N-terminal part of the I-TASSER model to generate a model consisting of four parallel helices linked by three turns (figure 21). This model was used as the starting structure for molecular dynamics simulations.

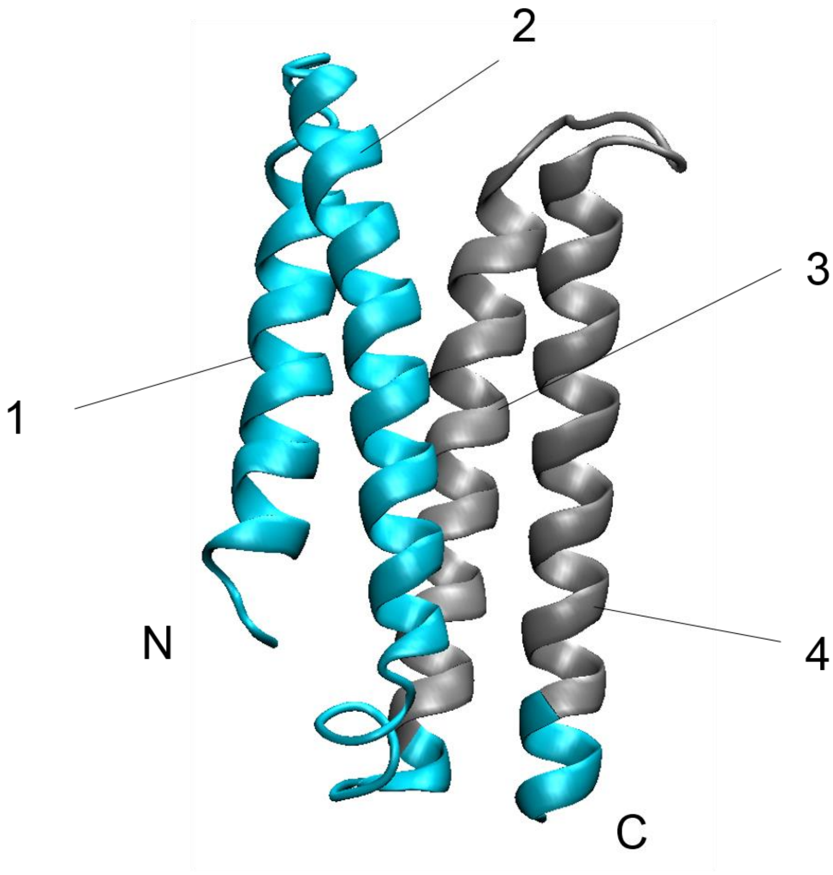


Figure 21. Model of the influenza A nuclear export protein with helices numbered. Regions modelled by I-TASSER are shown in blue.

From three 100 ns simulation trajectories and cluster analysis, it was observed that the conformation of the N-terminal (1-65) did not change significantly from the initial model; the experimentally known parts of the structure were kept restrained to their positions during the simulation. The backbone RMSD for the unrestrained part of the NEP from each simulation is shown in figure 22 for each simulation. Based on the RMSD plots, the trajectories were analysed to identify clusters of structures sampled starting from 10 ns with an RMSD cut-off of 0.2 nm. Simulation one produced eight clusters, simulation two produced nine clusters and simulation three produced ten clusters. The central structures from cluster one, which was the largest cluster from each simulation were within 3.0 Å of each other. The RMSF of N-terminal residues was also analysed (figure 23) and the most flexible residues were found to be between positions 22-27 and 53-57.

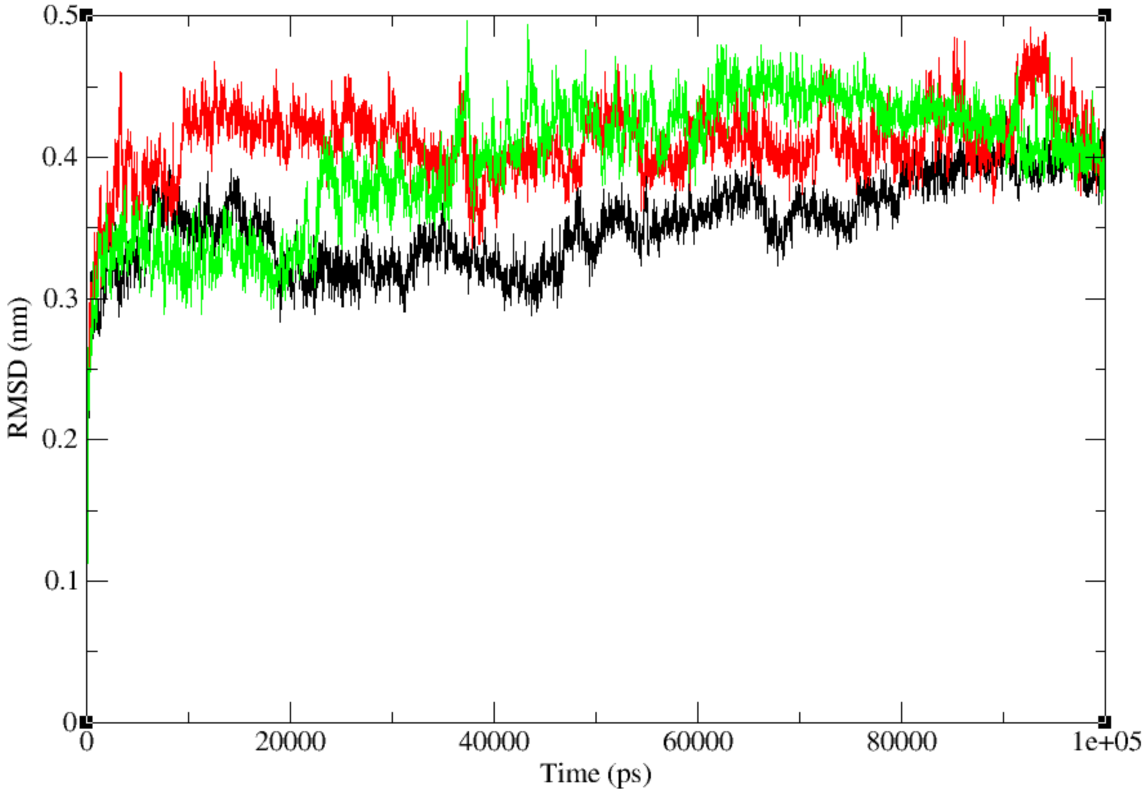


Figure 22. Root mean square deviation of NEP backbone atoms of residues 1-65 for three replicate 100 ns simulations, relative to the starting structure.

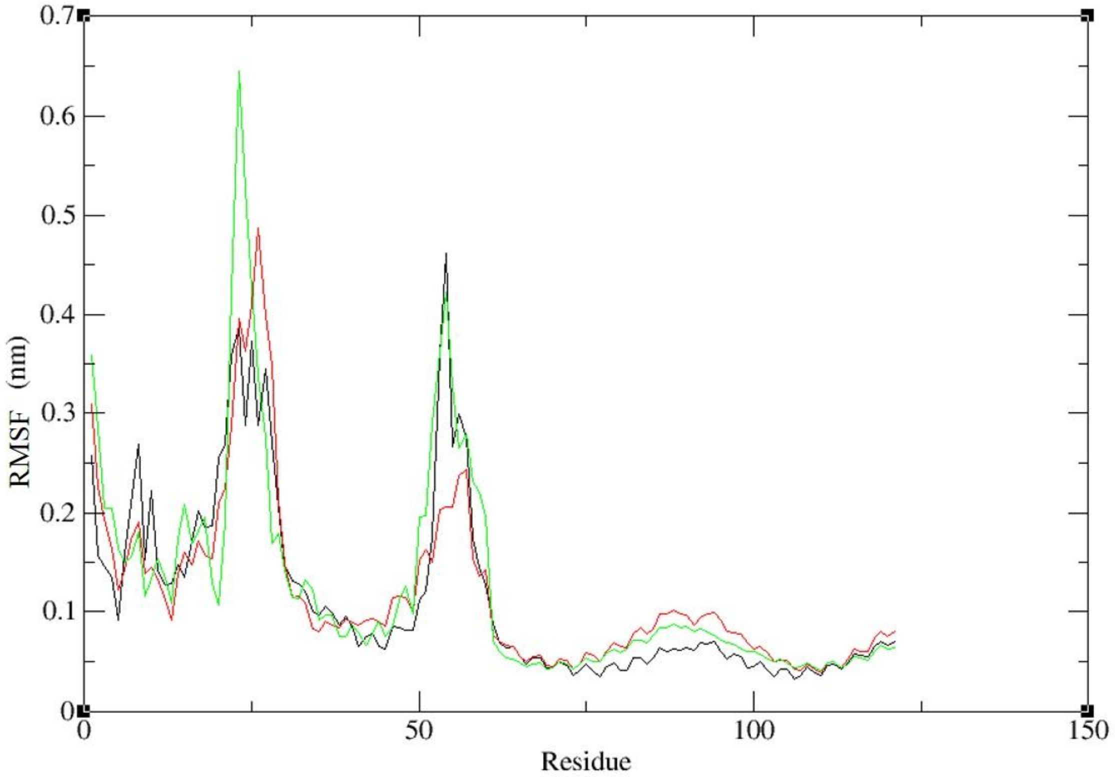


Figure 23. Root mean square fluctuation of NEP residues during three 100 ns simulations.

4.3.3 Intrinsic disorder prediction

Disorder prediction was performed using DISOPRED3. The intrinsic disorder profile (figure 24) shows that the predicted NEP structure based on the H1N1 sequence is unlikely to be disordered. The N-terminal half shows more disorder than the C-terminal half and possible disorder is indicated at residues Met1 and Asp2. A protein binding site was predicted within the disordered region indicating possible disorder to order transition upon protein binding.

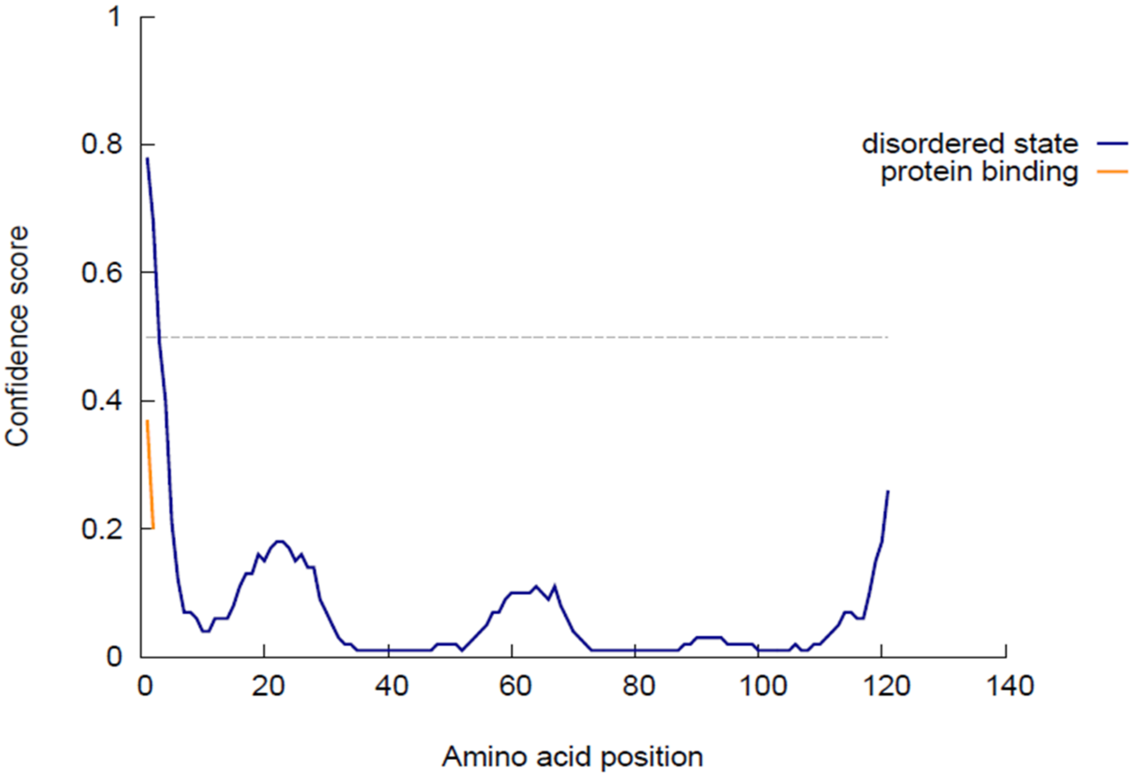


Figure 24. Intrinsic disorder profile for the nuclear export protein predicted from the sequence A/Puerto Rico/8/1934.

4.3.4 Computational solvent mapping

The representative structures from the largest cluster from each simulation were uploaded to FTMap to identify potential binding hot spot locations. Common hot spots in the hydrophobic region between helix two and four were predicted for all three structures, although with different ranks assigned. The level of residue conservation together with hot spot locations is shown in figures 25-27. A distance of 4.0 Å was chosen to indicate the amino acid residues surrounding the top five ligand binding spots (table 9-11).

Ten hot spots were identified in different regions on the structure from simulation one (figure 25) with three hot spots (ranked first, second and fourth) located between helices two and four. Three spots were predicted between helix one and two. No spots were located in the loop regions connecting helices one and two or helices three and four. All spots are surface exposed and the top five are closely surrounded by residues of high or intermediate conservation. Spot one, two and four bind with the highest number of different FTMap probe types.

Ten hot spots were identified on the structure from simulation two (figure 26). Five spots were located in the space between helix two and four, including spots eight and nine, and three and ten which were clustered closely together. All hot spots are surface exposed. Hot spot six is in the loop region between helix one and two and is surrounded by residues at position 20-29 that display high fluctuation. This spot is also very close to the highly variable residue Gly26.

Nine hot spots were identified on the structure from simulation three, all of which are surface accessible (figure 27). Spots one and four are located in the space between helix two and four, and are closely surrounded by highly conserved residues. Spot two is between helix one and three on the opposite side of the protein and close to Leu14 which is highly variable. No spots were located in the loop regions that join helix one and two, and three and four. The location of hot spot two from structure one was common amongst all three structures, and in a conserved region therefore was selected as the target site for virtual screening.

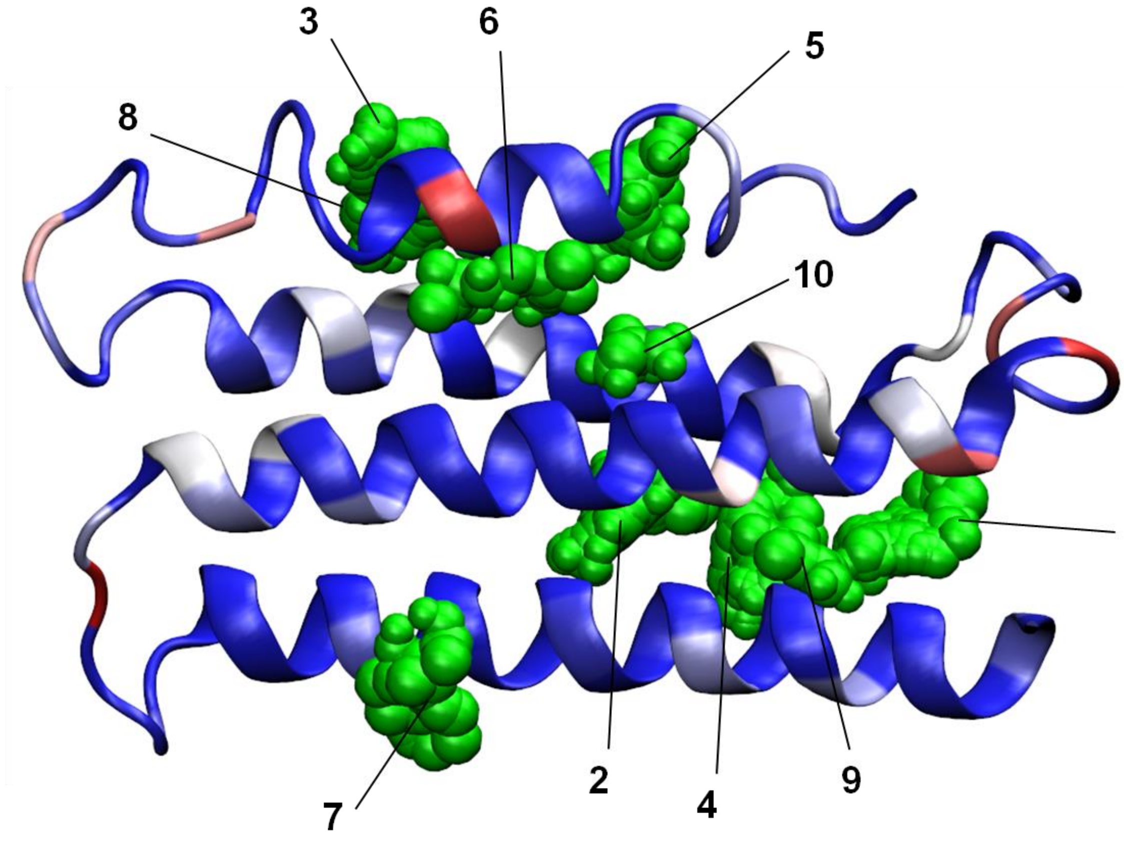


Figure 25. Ligand binding hot spots (green) identified by FTMap shown together with the degree of NEP conservation for the structure from simulation 1. The numbers indicate the site rank assigned by the algorithm.

Table 9. NEP amino acid residues within 4.0 Å of the five highest ranked binding hot spots predicted by FTMap from simulation 1. Highly conserved residues are shown in bold face.

Site	No. of probes bound	Residues within 4.0 Å of binding site
1	15	Met52, Val49, **Phe116**, **Leu120**, **Asn62**, **Ile113**, **Ser117**, **Arg66, Trp65**
2	15	**Arg42**, **Asp43**, **Gly46**, **Phe73**, **Val109**, **Leu105**, **Leu106**
3	13	**Gln20**, **Met19**, **Arg15**, **Ile12**, **Leu21**, **Tyr41**, **Met16**, Ser37
4	15	**Val109**, **Met50**, **Gly46**, Val49, **Phe116**, **Glu112**, **Ile113**
5	12	**Asn4**, **Thr5**, Ser7, **Ser8**, **Phe9**, **Ile12**, **Tyr41**, Leu40, **Ser44**

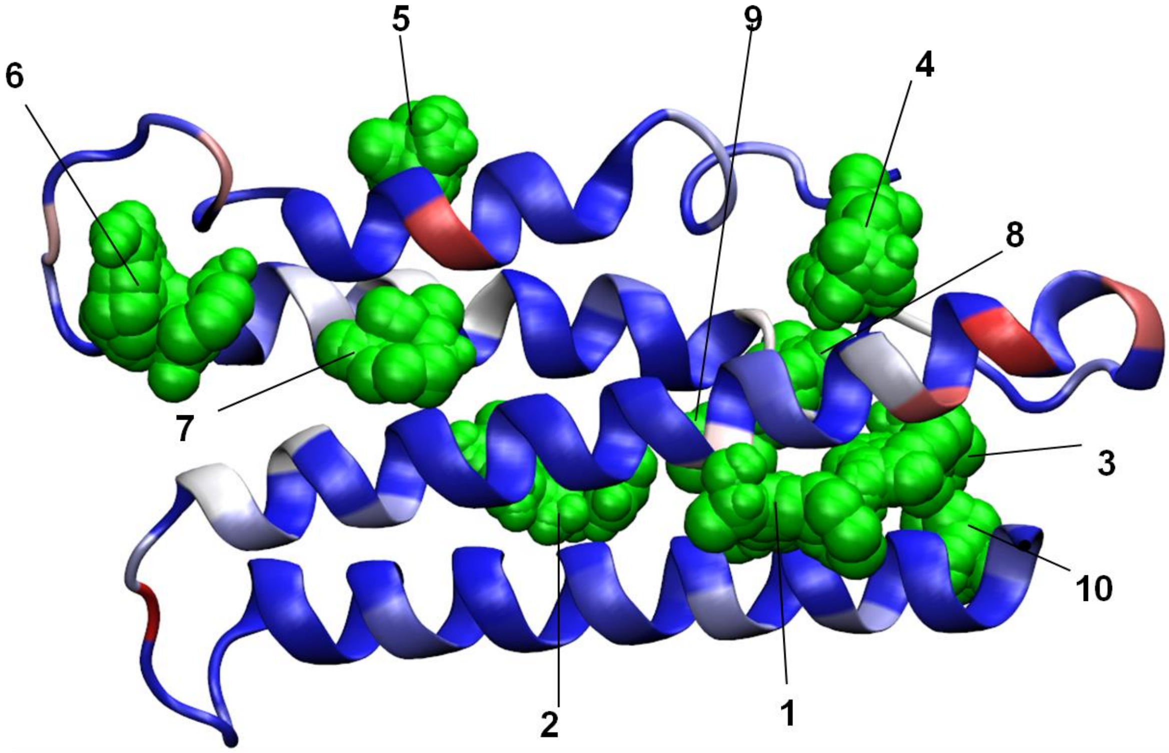


Figure 26. Ligand binding hot spots (green) identified by FTMap shown together with the degree of NEP conservation for the structure from simulation 2. The numbers indicate the site rank assigned by the algorithm.

Table 10. NEP amino acid residues within 4.0 Å of the five highest ranked binding hot spots predicted by FTMap from simulation 2. Highly conserved residues are shown in bold face.

Site	Number of probes bound	Residues within 4.0 Å of binding site
1	13	**Arg114**, **Glu67**, **Ser117**, **Glu110**, Glu63, **Gln71**, **Arg66**, **Ile113**, Gly70
2	13	**Lys39**, **Arg42**, **Leu105**, **Leu106**, **Ala102**, **Val109**, **Ile80**
3	11	Val49, **Met50**, Met52, **Gly53**, **Phe116**, **Ser117**, **Leu120**, **Arg66**, **Ile113**, **Trp65**
4	12	**Met1**, **Gln68**, Lys64, **Arg61**, **Val6**, Met52
5	10	**Met19**, **Tyr41**, **Arg15**, **Ile12**, **Met16**

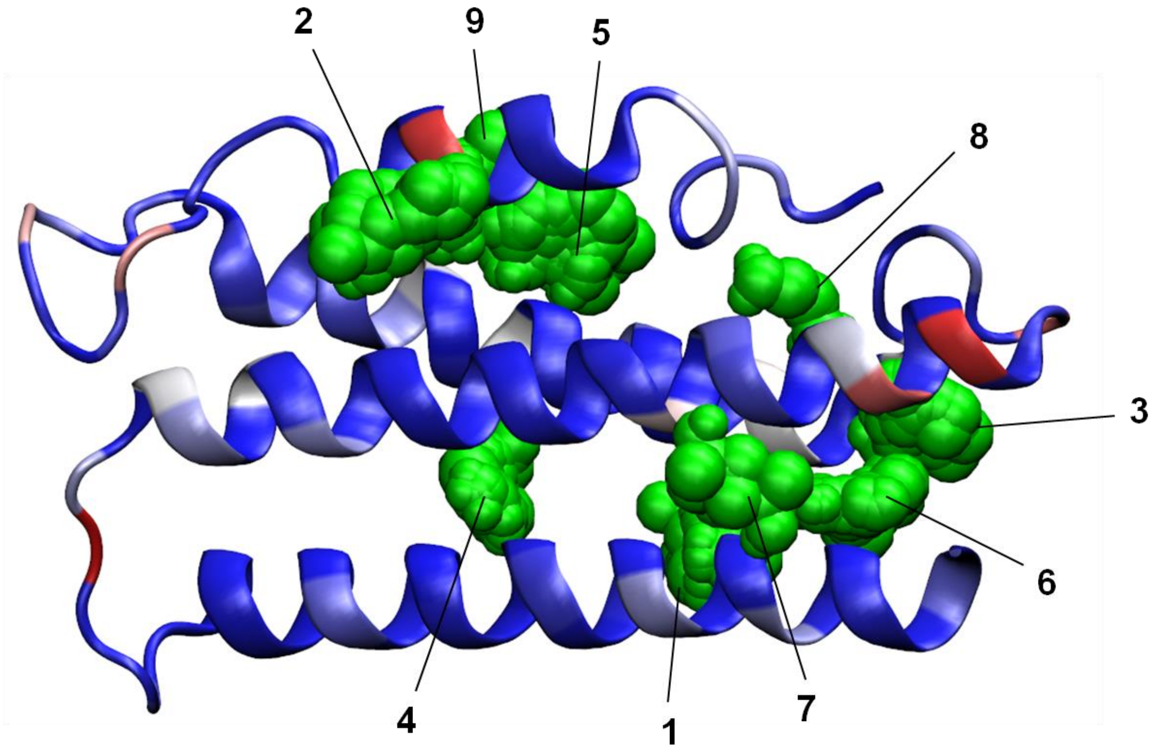


Figure 27. Ligand binding hot spots (green) identified by FTMap shown together with the degree of NEP conservation for the structure from simulation 3. The numbers indicate the site rank assigned by the algorithm.

Table 11. NEP amino acid residues within 4.0 Å of the five highest ranked binding hot spots predicted by FTMap from simulation 3. Highly conserved residues are shown in bold face.

Site	Number of probes bound	Residues within 4.0 Å of binding site
1	15	**Arg42**, **Val109**, **Gly46**, Val49, **Met50**, **Glu47**, **Phe116**, **Glu112**, **Ile113**
2	16	Leu14, **Glu82**, **Trp78**, **Ser17**, **Gln10**, **Glu75**, **Leu13**, **Leu79**
3	14	**Met50**, Met52, **Asn62**, **His56**, **Trp65**, **Gly53**
4	12	**Lys39**, **Arg42**, **Asp43**, **Ala102**, **Leu103**, **Leu105**, **Leu106**, **Ile80**
5	11	Leu40, **Ile12**, **Ser44**, **Tyr41**, Ser37

4.3.5 Virtual screening

The predicted binding affinities for 42,348 molecules from the NCI library ranged from -8.95 kcal/mol to +20.63 kcal/mol using AutoDock 4 (figure 28). The majority of compounds were found to bind within the range of -4.0 kcal/mol and -5.0 kcal/mol and 46 compounds had a positive binding energy score. The predicted binding affinities ranged from -8.7 kcal/mol to +34.9 kcal/mol using AutoDock Vina (figure 29). The majority of compounds were found to bind within the range of -4.5 kcal/mol and -5.5 kcal/mol and 17 compounds had a positive binding energy score. Several of the same compounds were identified with positive scores using both software.

The AutoDock Vina and AutoDock 4 rank lists were combined and the properties of ten top compounds from the consensus rank list are presented in table 12. Common amino acid residues involved in binding some of these compounds through hydrogen bonding and hydrophobic interactions identified using LigPlot+ include Arg42, Asp43, Lys39, Ile80, Gln101, Leu105, Val109; all of which are highly conserved. The compound ZINC01717023 was identified as a top hit molecule with the most similar binding scores from both software. The docking model of compound ZINC01509994 in the NEP target site, as well as predicted interactions is shown in figure 32.

1738 compounds from the DrugBank-approved library were also screened against the same NEP target site. The binding affinities ranged from -8.31 kcal/mol to +22.59 kcal/mol using AutoDock 4 (figure 30) and from -7.7 kcal/mol to +7.4 kcal/mol using AutoDock Vina (figure 31). 29 drugs had a positive binding score with Autodock 4 compared to 2 drugs with AutoDock Vina (not included in the binding score distribution graphs). None of the approved drugs had a significantly stronger predicted binding affinity than the top ranked compound of the NCI library. The top ranked drug from the consensus rank list was found to be the steroid nandrolone phenylpropionate (ZINC03881613).

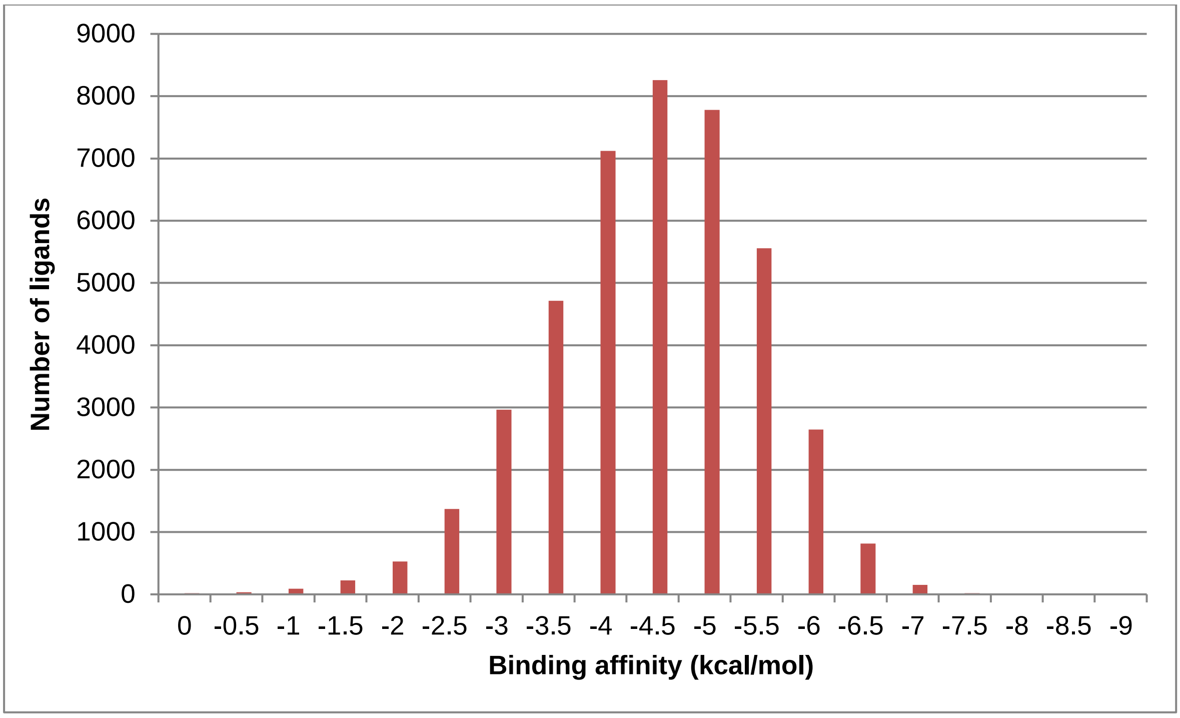


Figure 28. Frequency distribution plot of the binding affinities for ligands screened using AutoDock 4 from the NCI library. Negative scores shown only.

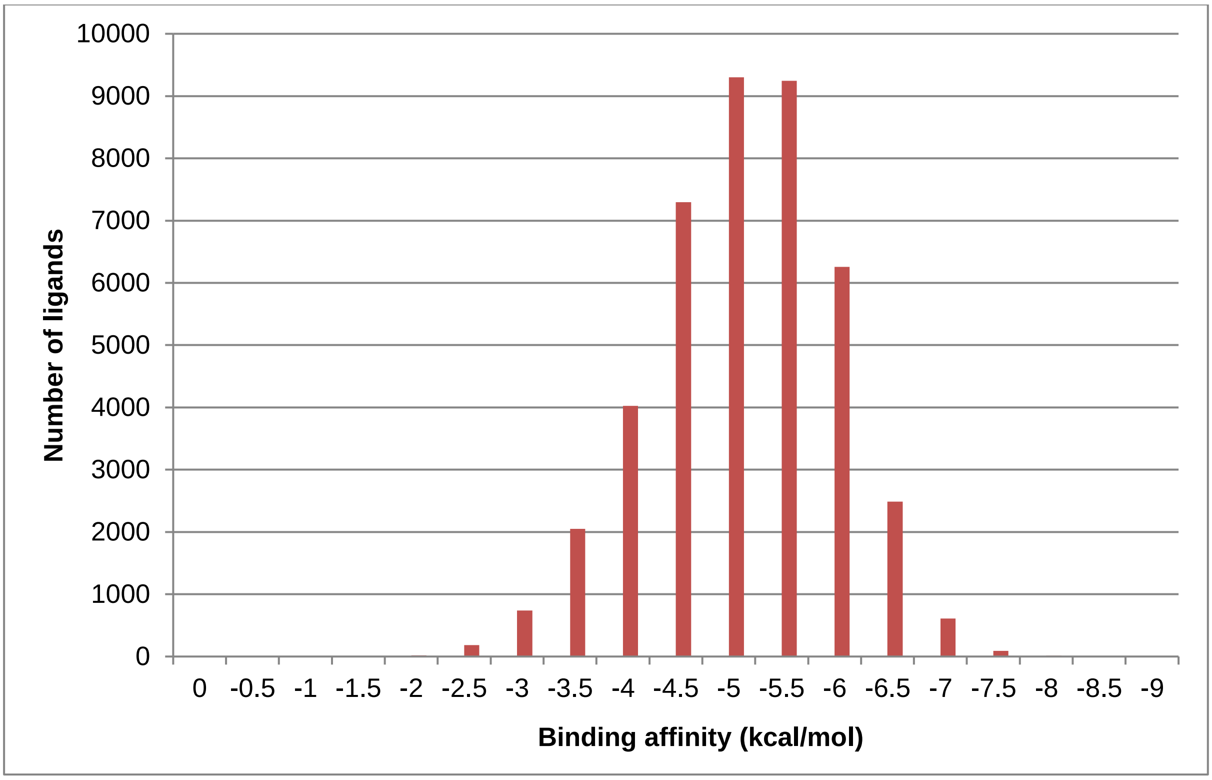


Figure 29. Frequency distribution plot of the binding affinities for all ligands screened using AutoDock Vina from the NCI library. Negative scores shown only.

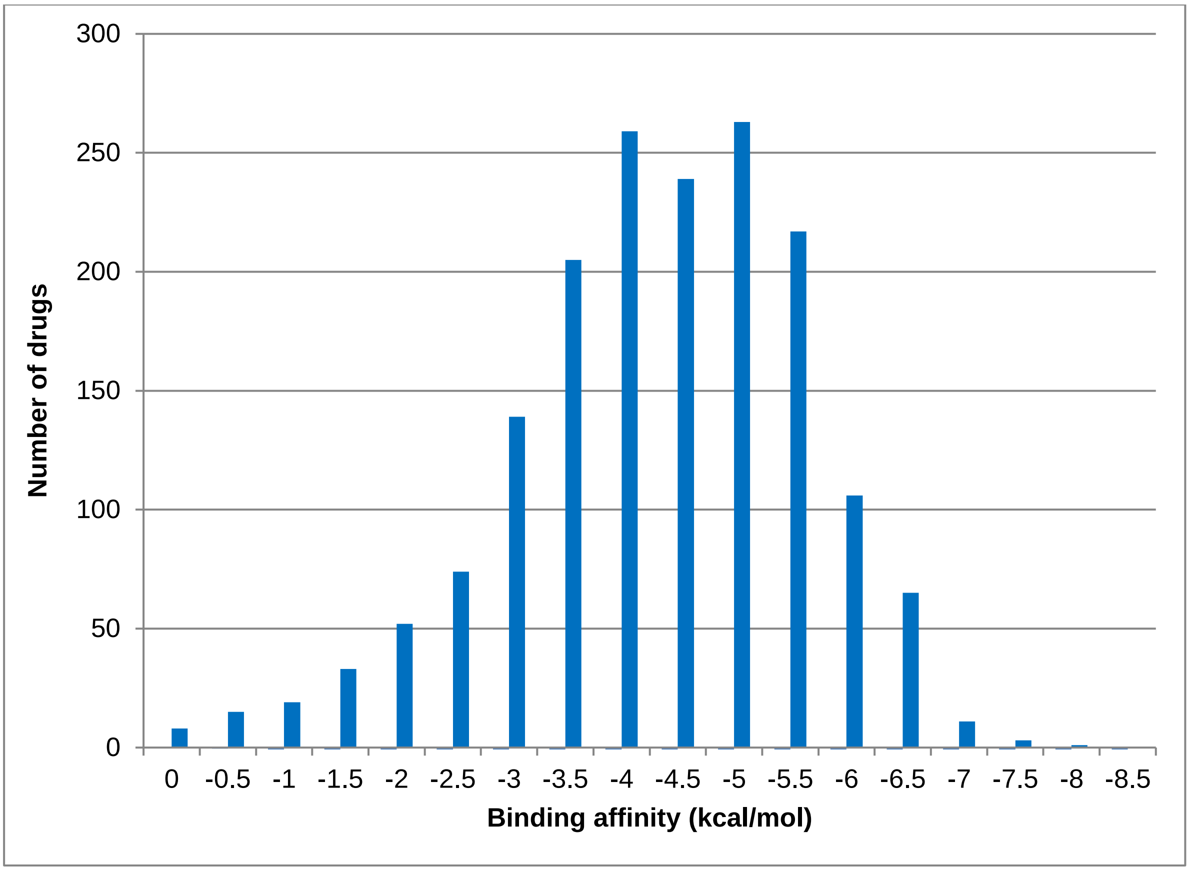


Figure 30. Frequency distribution plot of the binding affinities for drugs from the DrugBank library screened using AutoDock 4.

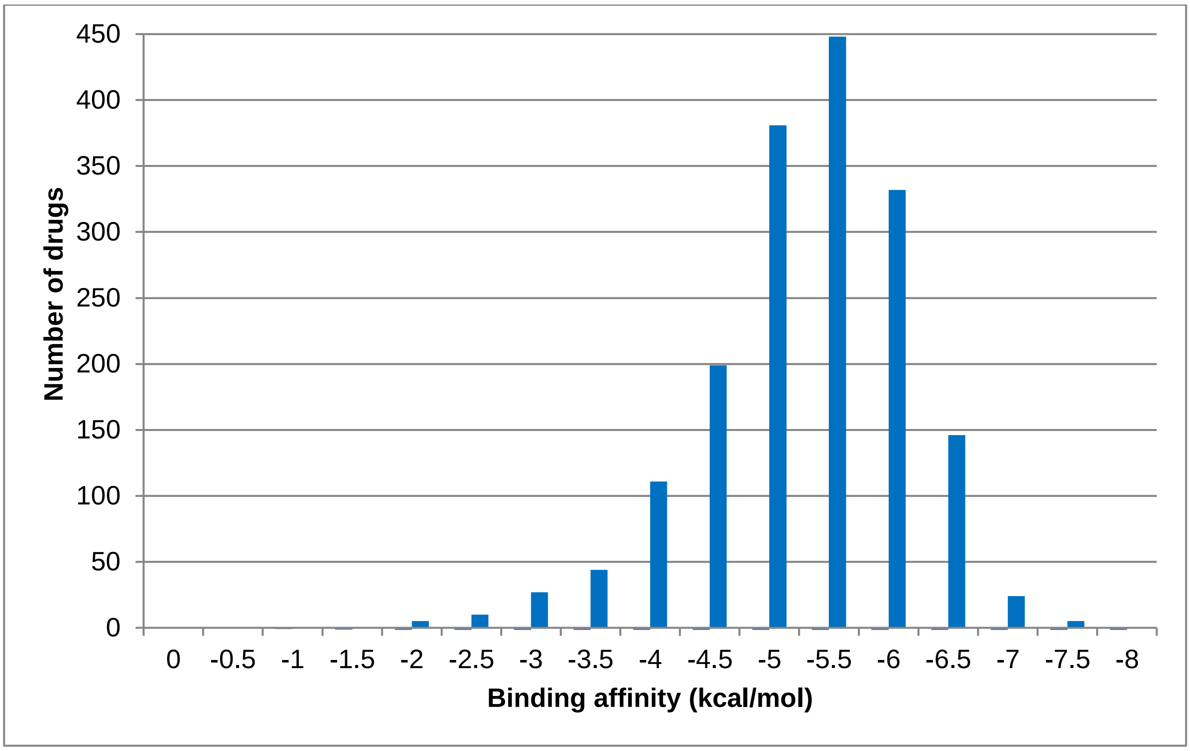


Figure 31. Frequency distribution plot of the binding affinities for drugs from the DrugBank library screened using AutoDock Vina.

Table 12. Properties of top compounds from the NCI and DrugBank libraries identified by consensus virtual screening based on binding affinity (ΔG), number of Hydrogen bond donors (H Don) and acceptors (H Acc), molecular weight and partition coefficent (xLogP) at pH 7 from the ZINC database.

Compound (ZINC ID)	**ΔG (kcal/mol)**		**xLogP**	**H Don**	**H Acc**	**Mol Wt (g/mol)**
	AD4	**Vina**				
NCI library						
ZINC01564229	-8.95	-7.4	-3.26	0	2	360.460
ZINC01717023	-8.09	-8.7	5.29	0	5	417.420
ZINC01509994	-8.52	-7.6	4.31	0	9	436.646
ZINC02035101	-6.9	-8.3	3.48	0	5	357.365
ZINC01592475	-8.10	-6.5	1.63	0	8	345.283
ZINC01646246	-7.34	-8.2	5.62	0	2	332.358
ZINC01717021	-7.07	-8.2	5.29	0	5	417.420
ZINC01561925	-8.02	-7.5	5.53	0	6	406.444
ZINC08617587	-7.93	-6.2	4.21	0	1	270.416
ZINC01645196	-7.83	-6.6	3.57	2	7	405.892
DrugBank library						
Nandrolone phenylpropionate	-6.83	-7.7	5.51	0	3	406.566
Estropipate	-8.3	-7.1	0.84	0	5	349.428
Pimozide	-4.56	-7.6	5.62	2	4	462.564
Tretinoin	-7.8	-6.0	5.80	0	2	299.434
Ergotamine tartrate	-6.28	-7.6	2.08	3	10	581.673
Adapalene	-7.5	-6.6	7.69	0	3	411.521
Azelastine Hydrochloride	-6.88	-7.5	4.82	1	4	382.915
Tasosartan	-7.50	-7.1	2.48	0	8	410.461
Azelastine hydrochloride	-6.84	-7.5	4.82	1	4	382.915
Zuclopenthixol	-7.33	-6.4	4.69	2	3	401.983

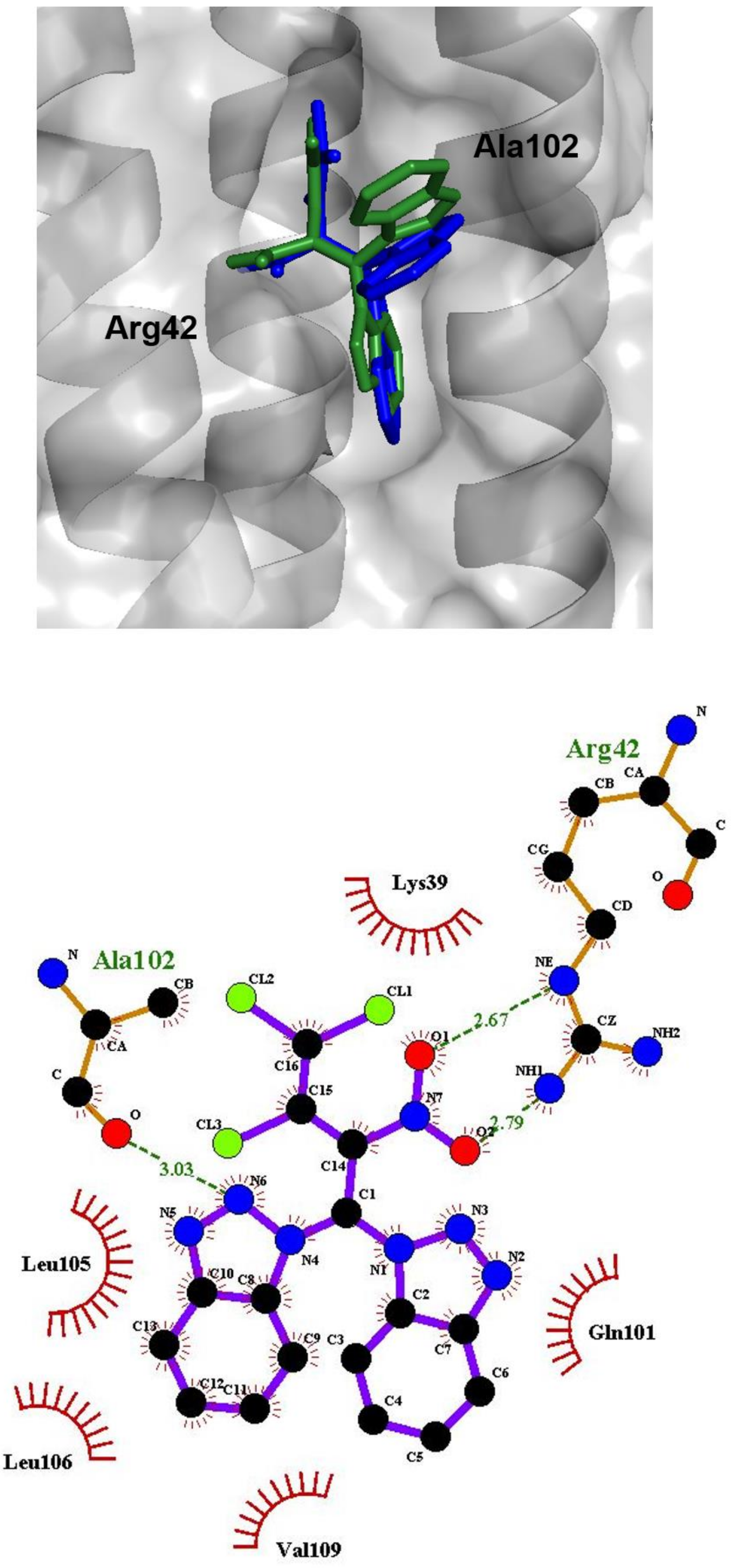


Figure 32. 3D docking model of compound ZINC01509994 in the NEP target site predicted with AutoDock 4 in blue and Vina in green (top) and 2D plot showing molecular interactions from the AutoDock 4 pose, green dash lines represent hydrogen bonds (bottom).

4.4 DISCUSSION

4.4.1 NEP conservation

Conservation results show that the influenza A nuclear export protein (NEP) is highly conserved as 112 out of 121 amino acid residues scored above 0.800 based on the Valdar conservation scores from the sequence alignment. The most conserved region with several residues scoring over 0.950 was the C-terminal helix four, which has been reported to stabilise the structure of helix three by interacting with the M1 protein through intermolecular bonds between side chains (Akarsu et al., 2003). Gly30, Leu38 and Arg66 were found to be 100% conserved and could therefore be most resistant to change due to evolutionary adaption of the virus. In comparison to previous research by Darapaneni, Prabhakar and Kukol (2009), the results presented here also show that the inclusion of many more protein sequences and the use of a different scoring method to quantify conservation do have an influence on conservation classification. This is seen with residues Ser23, Met31, Lys39, Asp47, Phe58, Glu67, Leu107, Phe116, and Phe118, which were reported as variable in the previous study, but were found to be highly conserved in this work. Despite these exceptions, the overall pattern of conservation is otherwise similar.

With regards to known NEP functions, the N-terminal region is reported to contain two leucine rich nuclear export signal (NES) motifs at position 11-21 and 22-45 (Huang et al., 2013). In this work, the second NES region was found to consist of four highly conserved leucine residues at position 21, 28, 38 and 45. This feature assists recognition and binding of cellular CRM1-RanGTP to the NES (Akarsu et al., 2003) and is a critical stage in transport of vRNPs out of the nucleus. Furthermore, a stretch of conserved hydrophobic residues between position 31-40 (Met31, Phe35, Leu38 and Leu40) of helix two have an effect on the nuclear and cytoplasmic distribution of NEP, as mutation of these residues to alanine resulted in NEP nuclear retention and also reduced virus growth (Huang et al., 2013). In another study, mutation of methionine at positions 14, 16 or 19 and leucine at position 21 which make up the first NES showed reduced growth of H1N1 virus in cell culture (Iwatsuki-Horimoto, Horimoto, Fujii, & Kawaoka, 2004).

The host protein AIMP2 which is a tumour suppressor was found to interact with NEP in the cytoplasm. AIMP2 binds at the NEP N-terminal to positively regulate virus replication by preventing degradation of the M1 protein required for vRNP export (Gao et al., 2015). Additionally, the cellular protein human nucleoporin 98 (hNup98) has been found to interact with NEP of H5N1 subtypes at the 22-53 region which causes NEP to co-localise in the nucleoli, but the significance of this interaction is unclear (Chen et al., 2010). Overall, these findings suggest that the high conservation of the N-terminal region could be attributable to these functions.

The viral M1 protein binding site on helix three of the C-terminal region involves Trp78 and surrounding glutamate residues which are also highly conserved. This region is recognized by a nuclear localization signal (NLS) on the M1 protein (Akarsu et al., 2003; Shimizu, Takizawa, Watanabe, Nagata, & Kobayashi, 2011) which in turn is bound to vRNP's to form the export complex. However, mutation of Trp78 was also reported to have no effect on vRNP export and NEP function (Robb et al., 2009). Other highly conserved residues may have a role in maintaining the structure of the N-terminus, and forming recognition sites for other biomolecules.

Regions of low conservation may indicate recognition sites by the host immune system. Previous conservation analysis of sequences for eleven influenza A proteins to identify conserved immunogenic peptides as vaccine targets predicted several potential T-cell epitopes, although sufficient sequence conservation for T-cell epitope predication was not identified for the NEP (Heiny et al., 2007). Upon antigen processing, the NEP induces a T-cell response, but unlike for HA and NA there are no reports of specific neutralising antibodies being produced. Therefore despite being conserved, if the protein is not sufficiently immunogenic or abundantly present, a NEP peptide based vaccine may induce very little, long term protective humoral immunity *in vivo*.

Sequence variations may also be a result of subtype or species specific codon changes which have circulated and consequently lowered the alignment conservation score. Alanine at position 48 which displays intermediate conservation with a score of 0.791 enables efficient interaction with CRM1 and has been shown to reduce nuclear aggregation of NEP in a H1N1 strain (Gao et al., 2014). Significant differences between sequences could also lead to changes in biological functions and enable the protein with additional features that could result in host adaptations. This is exemplified by a compensatory mutation found in the NEP sequence of a human H5N1 isolate at position 16 (M16I) which originated in birds and was found to enhance the activity of avian derived polymerases in human cultured cells (Mänz, Brunotte, Reuther, & Schwemmle, 2012).

4.4.2 Structure and predicted binding site locations of the NEP

To investigate binding hot spots as drug target sites in conserved regions, a full length NEP structure is required, which was not available from the PDB, hence protein structure prediction and molecular dynamics simulations were performed. Contrary to an experimental study focusing on NEP structure, the simulation trajectories and RMSD based cluster analyses reveal that the NEP is a stable structure and that the N-terminal does not display significant conformational flexibility; a feature which is proposed to assist in formation of the nuclear export complex (Lommer & Luo, 2002).

Compared to the initial starting structure, helix one becomes slightly distorted throughout the 100 ns simulations with the end of the helix unravelling, whereas the structure of helix two mostly remains intact. The distances between the apex of the N-terminal and C-terminal helices also varies a little between simulations, presumably due to higher flexibility of the residues that form the loop connecting helix one and two, opening up a pocket at one end of the protein. Otherwise, no major structural changes occur that alter the predicted four helix bundle structure. Simple conventional MD simulations of nanosecond time scale provide an insight into structural dynamics of proteins, although, they do not provide complete conformational sampling, therefore more advanced methods would be required. Such methods developed for this purpose include 1) replica exchange MD, where independent parallel simulations (replicas) are run at different temperatures and are periodically exchanged depending on temperature and potential energy differences, 2) accelerated MD, where a bias potential is added to raise the potential energy surface near the minima and accelerate simulation time and 3) metadynamics; a tool which broadly explores the free-energy landscape of biomolecules (Bernardi, Melo, & Schulten, 2015; Gedeon, Thomas, & Madura, 2015).

The disorder prediction (DISOPRED3) result shows that the NEP structure is not intrinsically disordered, which supports the MD simulation results. Although, the low confidence score may be because of a low match between any potentially disordered regions of the NEP amino acid sequence against sequences of proteins that have missing regions of electron density in their PDB structures (Jones & Cozzetto, 2015). Despite cluster structures from each simulation being similar based on the protein backbone conformation, different binding hot spots were predicted for the three largest cluster structures by the FTMap server. This was most likely caused by different orientations of amino acid side chains such as those at positions 21-30, which permit different interactions with molecular probes. However, between the three structures, many binding hot spots were predicted in similar locations and close to the same residues. Most hot spots were located in conserved areas of the protein and several sites were predicted close together. Two hot spots were identified near the first nuclear export signal region (residues 11-23), either side of helix one on all three structures. The specific NEP-CRM1 binding location which has also been shown to be independent of the nuclear export signals (Huang et al., 2013; Neumann et al., 2000), may involve N-terminal residues on the outer face of the protein which form these hot spots. One hot spot on all structures was identified in close spatial proximity to residues 74-82 of helix three, which forms the viral M1 protein binding site.

Several other NEP binding proteins have been discovered, such as ATP5E, HINT2, SMC3 and F1F0-ATPase; however, their binding site locations are unknown (Gao et al., 2015; Gorai et al., 2012). As interaction with these proteins may have indirect effects on influenza replication, it is possible that their binding locations may coincide with predicted hot spots identified in this work. Those hot spots which are the same across the three structures are presumably important interaction sites with other biomolecules, as they are unaffected by different conformation states. As common hot spots in the hydrophobic region between helix two and four were predicted for all three structures, and consisted of several highly conserved residues, this part of the protein was selected as the target site for screening. This region was also identified to interact with several different FTMap probes.

4.4.3 Virtual screening

From the NCI library, compounds such as ZINC01717023 and ZINC01561925, which have comparable high binding affinities and similar docking poses predicted using both AutoDock Vina and AutoDock 4 could be the best candidate inhibitors. All of the top compounds from the combined rank list contain aromatic rings but their overall structures are very diverse. The predicted binding modes show that many of the top compounds bend around a cleft on the surface of the protein. Specific molecular interactions were analysed using the software LigPlot+, as shown in figure 32. The compound ZINC0150994 specifically interacts with Arg42 on helix two and Ala102 on helix four via hydrogen bonding, as well as Leu106, Leu105, Val109, Gln101 and Lys39 through hydrophobic interactions.

The top compound from the DrugBank library (Nandrolone phenylpropionate) is held in place between helices two and four through hydrophobic interactions only, with several surrounding residues of the target site. The DrugCard entry (DB00984) in the DrugBank database (Law et al., 2014; Wishart et al., 2006) states that this drug targets androgen receptors and can be used to treat haematological disorders, growth failure and Turners syndrome. Binding of larger compounds such as ZINC01561925 could partly block recognition of the second nuclear export signal on helix two, or accessibility of other binding proteins.

The overall binding affinity distribution for molecules from the NCI library and DrugBank library are quite similar between AutoDock 4 and AutoDock Vina, as shown in the graphs in figures 28-31. The bulk of binding affinities could be represented by molecules/drugs which have a common scaffold that display the same molecular interactions and therefore bind with an affinity of ~-5.0 kcal/mol to the target site. However, the distribution amongst outliers is less comparable between the software for the NCI library, as 716 molecules were identified with a binding affinity between -7.0 kcal/mol and -9.0 kcal/mol with AutoDock Vina, compared to 176 with AutoDock 4. These molecules may have structural features more specific to the target site, although many of these could be false positive predictions. Also, for both libraries a higher number of molecules with scores between -3.0 kcal/mol and 0 kcal/mol screened with AutoDock 4 are outliers from the bulk of binding affinities compared with AutoDock Vina.

5 THE POLYMERASE BASIC PROTEIN 2 (PB2)

5.1 INTRODUCTION

This chapter describes the work published in Patel & Kukol, (2017). The polymerase basic protein 2 (PB2) is encoded by RNA segment one. It is one of the largest influenza proteins consisting of 759 amino acids and is a constituent subunit of the trimeric viral polymerase complex. During transcription of the viral genome, the PB2 protein is mainly responsible for generating the cap structure for viral mRNA's from the 5' end of 7-methyl guanosine triphosphate (mGTP) capped host mRNA strand. The PB2 'cap snatching' mechanism involves residues between positions 318-482 which recognise methylated guanosine in order to bind the host cell RNA strand. The endonuclease subunit of the PA then cleaves the RNA, leaving a 10-13 nucleotide primer to initiate transcription by PB1 (Fodor, 2013). In complex, the N-terminal residues of the PB2 subunit are associated with the C-terminal subunit of PB1, which is a critical interaction to trigger acid polymerase (PA) endonuclease activity (Sugiyama et al., 2009). A detailed diagram of the PB2 subunit is presented in figure 33.

A structural study of the PB2 protein from a H5N1 avian virus had found that following translation, the C-terminal domain (residues 536-759) undergoes large conformational re-organisation between open and closed states. This flexibility enables the nuclear localization signal (NLS) peptide in the 686-759 region to bind with host importin-α, enabling PB2 entry into the nucleus of the target cell to catalyse further RNA transcription (Das, Aramini, Ma, Krug, & Arnold, 2010; Delaforge et al., 2015). Other PB2 conformational changes occurring in connection with the cap-snatching mechanism and the kind of RNA bound have also been described in the context of the full polymerase complex (Reich et al., 2014; Thierry et al., 2016).

In addition to mutations in the HA and NA proteins, changes in the sequences of polymerase proteins are considered as major determinants of host range and adaptation (Mehle & Doudna, 2009; Neumann & Kawaoka, 2015). The characteristic PB2 host determining residue at position 627 (with lysine being prevalent in human strains, glutamate present in avian strains and serine in bat strains) is situated in a loop region, which along with the cap-binding domain, does not make extensive contact with the PB1 and PA subunits (Kuzuhara et al., 2009; Pflug, Guilligay, Reich, & Cusack, 2014). The C-terminal 535-684 domain has also been shown to have RNA binding activity which is affected by the E627K mutation and is unrelated to the cap-snatching function (Kuzuhara et al., 2009).

As well as localising in the nucleus, the PB2 subunit interacts with mitochondrial antiviral signalling protein (MAVS) (Graef et al., 2010); a feature unrelated to genome replication. PB2 proteins of seasonal human strains have been found to be present in the mitochondria, whilst PB2 proteins of avian strains (such as H5N1) were not. This difference in localization is caused by a single amino acid polymorphism in the mitochondrial targeting signal which has been mapped to the N-terminal region at position nine (N9D). The significance of this finding is that non-mitochondrial associated PB2 variants induce higher levels of IFN-β in vitro, suggesting that PB2 also plays an important role in determining virulence by altering the immune response (Graef et al., 2010).

As the PB2 protein plays multiple essential roles in the virus life cycle, it is a valid target for antiviral drugs. Several crystal structures are available in the PDB for specific PB2 subunit domains in holo and apo forms which can aid with structure based drug discovery studies. Despite some of the PB2 surface area being inaccessible due to trimer assembly, inhibitors of the ‘cap snatching’ function to prevent capped host mRNA binding have been identified from cell based assays (figure 34) that have shown potent effects against several influenza strains *in vitro* (Boyd et al., 2015; Clark et al., 2014; Pautus et al., 2013). One of these compounds (VX-787, figure 34a) is currently undergoing clinical trials (Koszalka, Tilmanis, & Hurt, 2017).

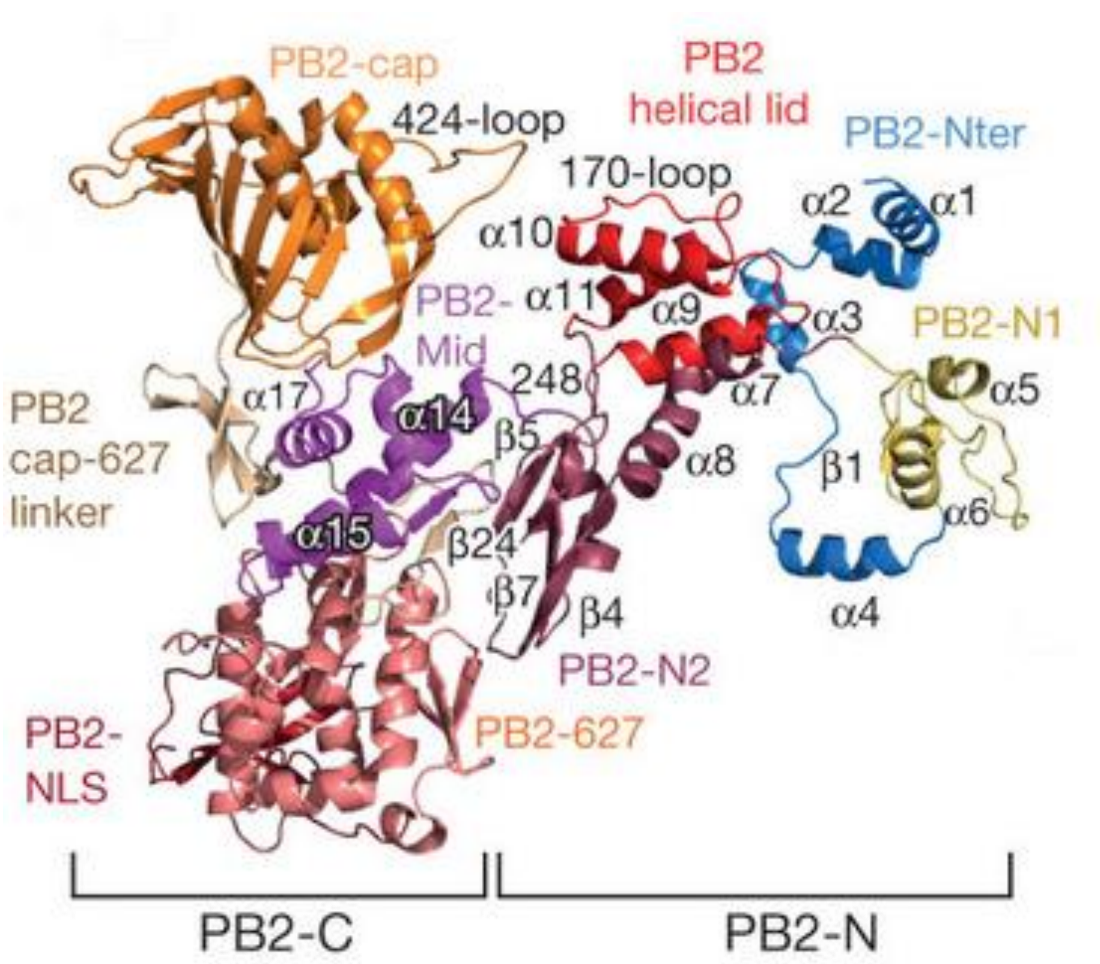


Figure 33. PB2 subunit showing the C-terminal third and N-terminal two thirds with secondary structure elements of the sub-domains coloured and labelled. Adapted by permission from Macmillan Publishers Ltd: Nature (Pflug et al., 2014), copyright (2014).

5.2 METHODS

5.2.1 PB2 sequence analysis

The sequence analysis was performed according to general methods section 2.2. Briefly, full length PB2 sequences were chosen from all hosts, regions and subtypes until January 2016. Identical sequences and sequences containing non-standard residues were removed. The CD-HIT web server (Huang et al., 2010) was used to cluster sequences meeting a similarity threshold of 98.5%, as this gave an acceptable number of sequences for further analysis. Global pairwise sequence alignment between the H5N1 (A/Vietnam/1203/2004) and H17N10 (A/little yellow-shouldered bat/Guatemala/060/2010) sequence was performed using the EMBOSS Needle tool version 6.6.0 (http://www.ebi.ac.uk/Tools/psa/emboss_needle/) with the default settings. For calculation of amino acid conservation see general methods section 2.3. The minimum score used for re-scaling was 0.789.

5.2.2 Protein modelling

The structure of a full length amino acid sequence of the PB2 polymerase isolated from a human host (A/Vietnam/1203/2004 (H5N1)) was predicted with the I-TASSER server (Yang & Zhang, 2015; Zhang, 2008). The I-TASSER model was aligned with the crystal structure of A/Vietnam/1203/2004 (PDB ID: 3L56) and the aligned coordinates were saved as a PDB file using PyMol. Experimentally known coordinates of residues 542-673 and 690-738 were copied from the crystal structure into the model. Missing atoms and side chains were inserted with Swiss-PDB Viewer version 4.1.0.

Energy minimization was performed in solvent with 1000 steps of the steepest descent algorithm to remove atomic clashes and to optimise torsion angles of the PB2 model. The PDB file was converted to a Gromacs file and molecular topology file with the program pdb2gmx. The AMBER99SB – ILDN force field was selected as it demonstrated high accuracy in a systematic evaluation of eleven molecular dynamics force fields (Beauchamp et al., 2012) along with the TIP3P water model. The protein was solvated in water in a cubic box with periodic boundary conditions, and the charge of the chemical system was neutralised with 21 chloride ions and additional NaCl at 100 mM concentration. The particle mesh ewald algorithm for calculating electrostatic interactions was used with a real space cut-off distance of 1.0 nm and the cut-off for van der Waals interactions was 1.0 nm. The following commands were used:

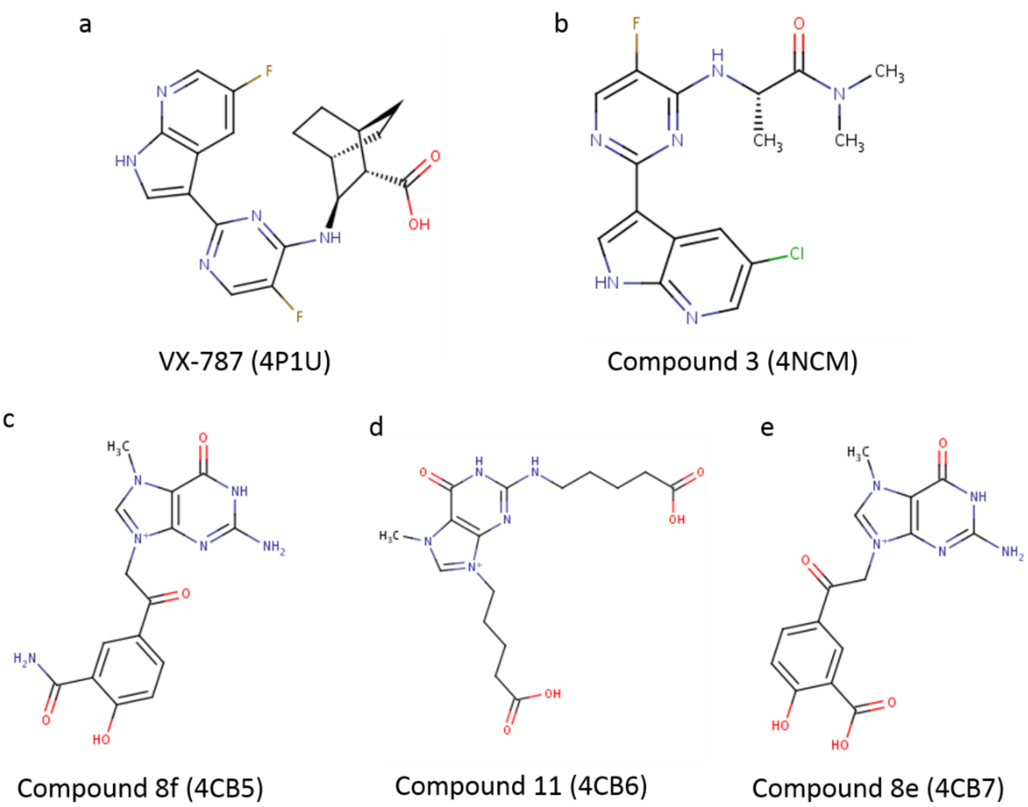
```
$ pdb2gmx -f PB2complete.pdb -o PB2complete.gro -p PB2complete.top -ter –ignh
$ editconf -f PB2complete.gro -d 0.75 -o box.gro -bt cubic
$ genbox -cp box.gro -cs spc216.gro -p PB2complete.top -o PB2solvated.gro
$ grompp -f ions.mdp -p PB2complete.top -c PB2solvated.gro -o ions.tpr
$ genion -s ions.tpr -neutral -conc 0.1 -p PB2complete.top -o ions.gro
$ grompp -f em.mdp -p PB2complete.top -c ions.gro -o em.tpr
$ mdrun -s em.tpr -c PB2after_em.gro
$ editconf –f PB2after_em.gro –o PB2after_em.pdb
```

5.2.3 Prediction of binding hot spots

The PB2 model after energy minimisation was uploaded to the FTMap server (Brenke et al., 2009). See general methods section 2.5.

5.2.4 Virtual screening - benchmarking

Five previously identified compounds which have shown to inhibit influenza replication *in vitro* and are reported to bind the PB2 polymerase (Clark et al., 2014; Pautus et al., 2013) were used to benchmark virtual screening methods to find the best method of identifying true positives at the top of the rank list. The structures of the active compounds and their PDB codes are shown in figure 34. These compounds were downloaded from the PDB and converted to the PDBQT format with the script ‘prepare_ligand4.py’, which is included in the AutoDock Tools package. The following parameters were specified: add hydrogen atoms and build bonds between any non-bonded atoms, merge both non-polar hydrogens and lone pairs, and allow peptide-backbone-bonds to rotate. The three methods tested were: AutoDock Vina version 1.1.1 (Trott & Olson, 2010), AutoDock 4 (Morris & Huey, 2009), and a consensus method using both (Kukol, 2011). 180 decoy molecules with similar molecular weight from the NCI Plated 2007 library were selected for the benchmarking. The grid parameters for the benchmarking were set around the mGTP binding pocket and remained the same as those reported in the publication by Pautus et al., (2013). The top ten positions of all 185 compounds ranked according to their binding affinity were considered for identifying the known inhibitors.



VX-787 (4P1U)

Compound 3 (4NCM)

Compound 8f (4CB5)

Compound 11 (4CB6)

Compound 8e (4CB7)

Figure 34. Chemical structures and PDB-IDs of five PB2 inhibitor compounds for benchmarking virtual screening methods.

5.2.5 Virtual screening against PB2 target site

A region encompassing binding hot spot two was selected as the target site for virtual screening. The grid box dimensions for docking were set using AutoDock Tools as follows with a grid spacing of 0.375 Å:

```
center_x = 58.942
center_y = 70.112
center_z = 63.114
size_x = 22.5
size_y = 20.25
size_z = 20.25
exhaustiveness = 12
```

The same compound library used for screening against the NS1 protein was filtered using the software Open Babel version 2.3.1 (O’Boyle et al., 2011) to eliminate compounds with molecular weight over 500 g/mol and predicted partition co-efficient (logP) over five using the command:

```
$ obabel output.mol2 --filter “MW<500 logP<5” –O filtered.mol2
```

The remaining 46,926 chemical compounds in .mol2 format were split into individual ligand files and saved in PDBQT format using the AutoDock screening preparation tool Raccoon.

Post screening, the library was converted to SDF format and passed through the Pan Assay Interference Compounds (PAINS) filter with the online FAF-Drugs3 (Free ADME-Tox Filtering Tool) program (Baell & Holloway, 2010; Lagorce et al., 2015) to identify and remove compounds from the rank list that appear as frequent hitters in screening experiments. A total of 42,348 compounds remained. The DrugBank-approved library containing 1738 drugs (Law et al., 2014) was downloaded from the ZINC database and also screened against the PB2 target site in order to find any approved drugs that may also target the PB2.

5.3 RESULTS

5.3.1 PB2 amino acid conservation

12,459 PB2 sequences were obtained from the NCBI influenza virus resource database. This included 31% from human, 16% from swine and 50% from avian hosts. 702 sequences remained after removing redundant sequences at 98.5% identity. The conservation scores calculated from the multiple sequence alignment of the non-redundant sequences shows that there is a high level of amino acid conservation throughout the entire protein sequence. The scores ranged from 0.789 (lowest) at position 147 to 1.0 (highest) and the majority of amino acids had a score between 0.95 and 1.0 (table 19, appendix 9.3). Overall, the key functional regions of the protein were found to be well conserved. 42 amino acids were found to be 100% conserved in the dataset. Low (scores below 0.850) or moderate conservation was identified mainly at single amino acid positions such as 64, 107, 147, 271, 292, 453, 483, 559, 588, 590, 591, 613, 661 and position 676 which were all located on the exterior surface of the protein. An intermediate level of conservation for the host specific residue at position 627 was reflected in the alignment with a conservation score of 0.885. For display purposes the conservation scores were re-scaled and mapped on to the PB2 structure (figure 35).

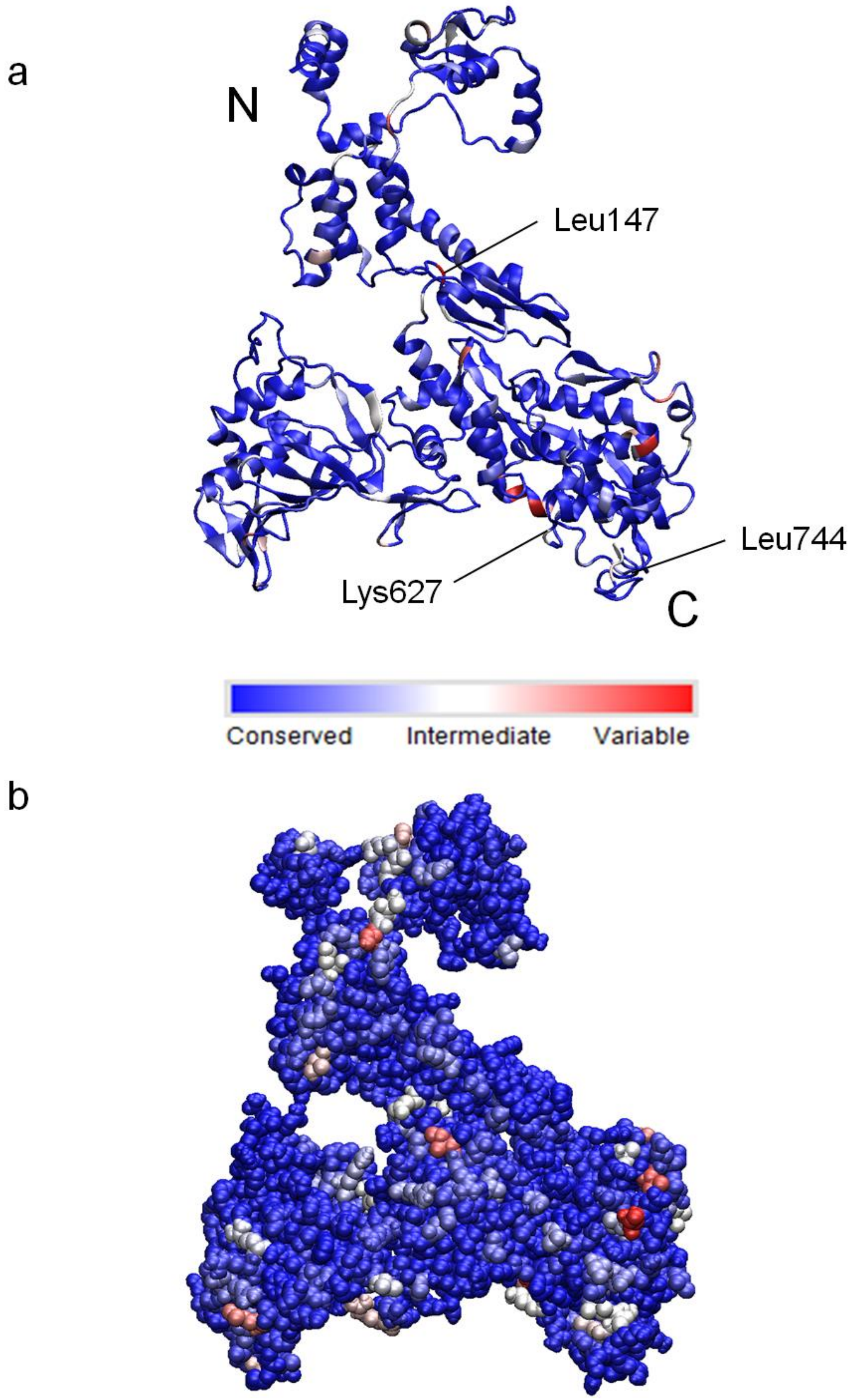
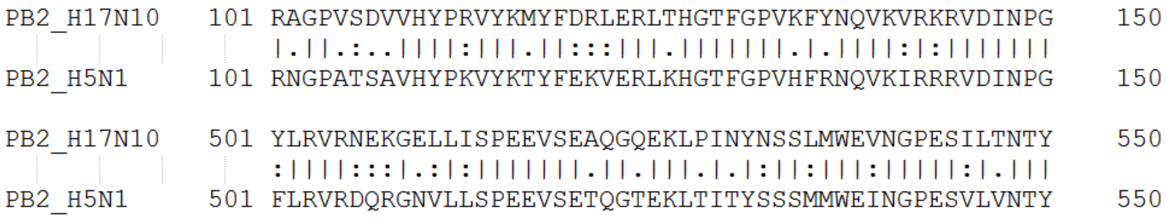


Figure 35. PB2 amino acid conservation mapped onto the H5N1 influenza A PB2 protein structure shown in (a) cartoon representation and (b) spacefill representation. Figure adapted from Patel & Kukol, (2017).

5.3.2 Protein modelling

The I-TASSER model of the human H5N1 sequence had a template modelling (TM) score of 0.92, and was largely built on the PDB template 4WSB chain C, which is from a H17N10 subtype of bat influenza A. Residues 483-490 and 742-759 which were not covered by any template were modelled *ab-initio*. The H5N1 PB2 fragment (3L56) is structurally very similar to that of H17N10 with a backbone RMSD of 1.05 Å. The final model was refined by energy minimisation. The overall percentage identity between the two full length sequences is 68% and the overall similarity is 83% calculated using the EBLOSUM62 matrix. The pairwise alignment of the two sequences covering some of the amino acid residues surrounding the target site selected for virtual screening is shown in figure 36. Many of the residues in the target site are located far apart in the amino acid sequence. Residues with an exposed surface area above 2.5 Å^2 were considered as exterior residues.



```
PB2_H17N10    101 RAGPVSDVVHYPRVYKMYFDRLERLTHGTFGPVKFYNQVKVRKRVDINPG    150
                  |.||.:..||||:|||.||:::|||.|||||||.|.||||:|:|||||||
PB2_H5N1      101 RNGPATSAVHYPKVYKTYFEKVERLKHGTFGPVHFRNQVKIRRRVDINPG    150

PB2_H17N10    501 YLRVRNEKGELLISPEEVSEAQGQEKLPINYNSSLMWEVNGPESILTNTY    550
                  :||||:::|.:|:|||||||.||.|||.|.|:||:|||:|||||:|.|||
PB2_H5N1      501 FLRVRDQRGNVLLSPEEVSETQGTEKLTITYSSSMMWEINGPESVLVNTY    550
```

Figure 36. PB2 H5N1 (A/Vietnam/1203/2004) and H17N10 (A/little yellow-shouldered bat/Guatemala/060/2010) pairwise sequence alignment of the region 101-150 and 501-550 covering the target site for virtual screening. A vertical line indicates identical residues, a colon indicates strong similarity and a period indicates weak similarity.

5.3.3 Computational solvent mapping

Fifteen ligand binding hot spots were predicted using FTMap in several domains, most of which were located in highly conserved areas of the protein. The top ten binding hot spots are annotated in figure 37. The most surface accessible spots are site three, four, five, six, seven, eight, nine, fourteen and fifteen. Hot spots two, eleven, twelve and ten appear to be partially buried when viewing the structure in spacefill representation, with sites one and thirteen being the least exposed to the outer surface on the protein core. The highest number of different probes were found to bind at site one. Site seven, fourteen and fifteen were clustered closely together in the mGTP capped RNA binding region, whereas site six was the only site located within the N-terminal third. Site three is close (within ~6.0 Å) to the intermediately conserved residue Val613. It was decided to consider the conserved region encompassing hot spot two as the target site for virtual screening.

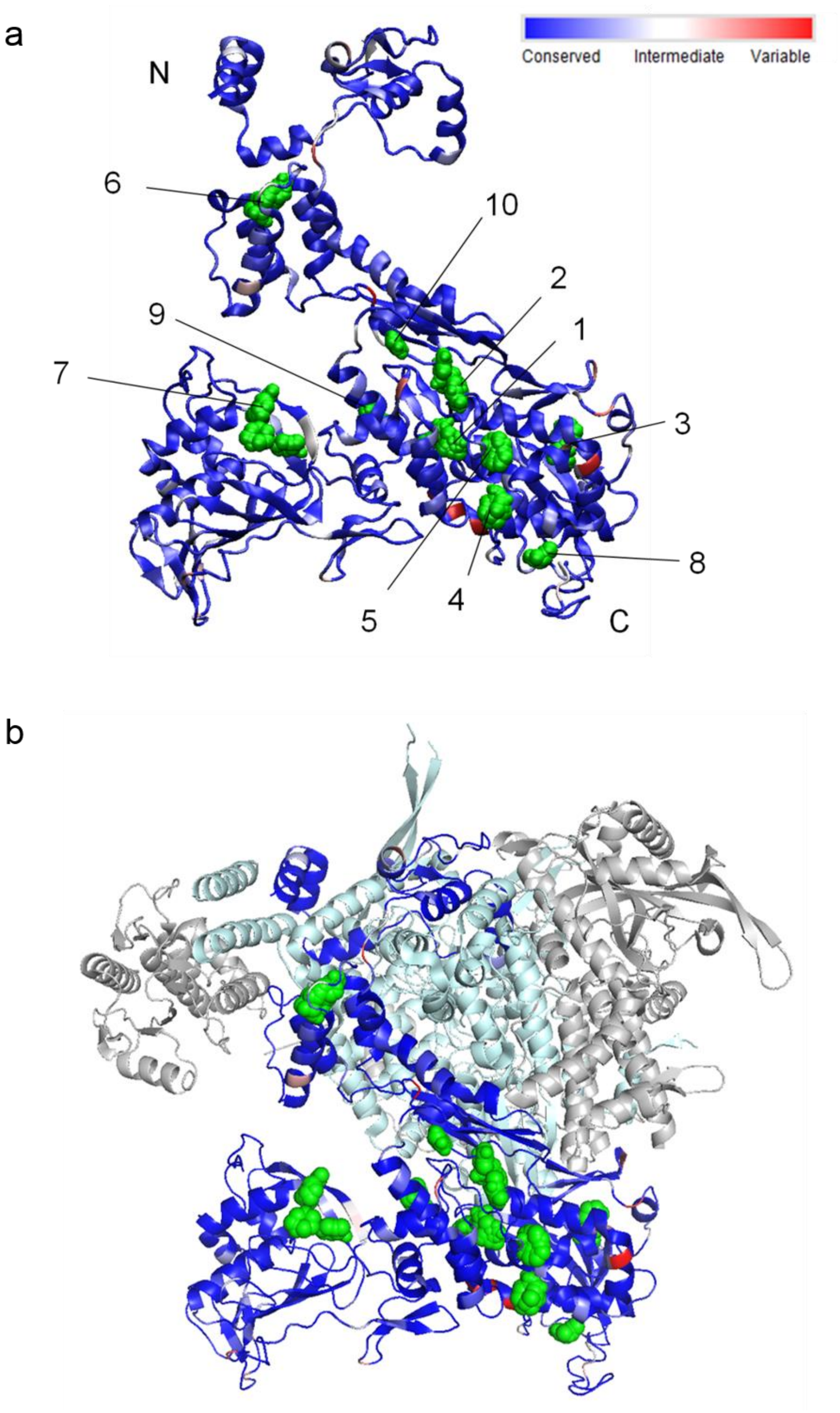


Figure 37. Locations of the top ten binding hot spots (green spheres) identified by the FTMap algorithm shown together with the degree of PB2 sequence conservation on (a) the H5N1 PB2 structure (Patel & Kukol, (2017)) and (b) the H5N1 PB2 subunit structurally aligned with the H17N10 polymerase complex. The numbers indicate the rank assigned by the FTMap algorithm.

5.3.4 Virtual screening - benchmarking

A virtual screening benchmark was performed against the mGTP capped RNA binding site. The ability of the docking software to identify five known inhibitors among the top ten predictions out of 180 compounds with similar molecular weight was tested. The results showed that AutoDock Vina alone and a combination with AutoDock 4 were the best method to retrieve active compounds within the top positions of the rank list as both methods were able to identify one inhibitor (table 13). The binding affinities using AutoDock Vina ranged from -7.4 kcal/mol to -4.5 kcal/mol and from -7.4 kcal/mol to -2.9 kcal/mol with AutoDock 4. For simplicity, the single AutoDock Vina software was used for the PB2 target site screening.

Table 13. Results from benchmarking three docking software for virtual screening.

Software	Number of true ligands found within top 10	Binding affinity of ligand (kcal/mol)
AutoDock Vina	1	-7.4
AutoDock 4	0	
AutoDock Vina + AutoDock 4	1	-7.4+ -6.0

5.3.5 Virtual screening - PB2 target site

The binding affinities of 46,926 compounds screened from the NCI library ranged from -10.3 kcal/mol to +13.7 kcal/mol using AutoDock Vina (figure 39). A large proportion of compounds were predicted to bind between the range of -5.0 kcal/mol and -7.0 kcal/mol. Fourteen ligands had binding energies above zero kcal/mol, (not shown on the energy distribution graph in figure 39) and several molecules at the top of the rank list had similar chemical structures. All compounds identified as PAINS were removed from the rank list to help selection of top hit compounds. Some of the key amino acid residues found to interact with the top ten compounds via hydrogen bonding and hydrophobic interactions include: Gln138, Gly222, Ile529, Ile539, Asn540, Gly541, Tyr531 and Thr530; all of which are highly conserved. The predicted binding conformations show that some of these compounds bind partially inside a deep pocket formed by these residues (figure 38).

The DrugBank-approved library containing 1738 small molecule drugs was also screened against the same target site. The binding scores ranged from -10.0 kcal/mol to +53.8 kcal/mol with the largest proportion of drugs predicted to bind between the range of -5.0 kcal/mol and -7.0 kcal/mol. 32 drugs had a positive binding score above zero kcal/mol (not shown on the energy distribution graph in figure 40). None of the approved drugs had a significantly stronger predicted binding affinity than the top ranked compound of the NCI library; however, the highest ranked drug paliperidone (ZINC04214700) had the same binding affinity (-10.0 kcal/mol) as three compounds of the NCI library ranked within the top ten positions. The chemical properties of top hit compounds from both libraries are shown in table 14.

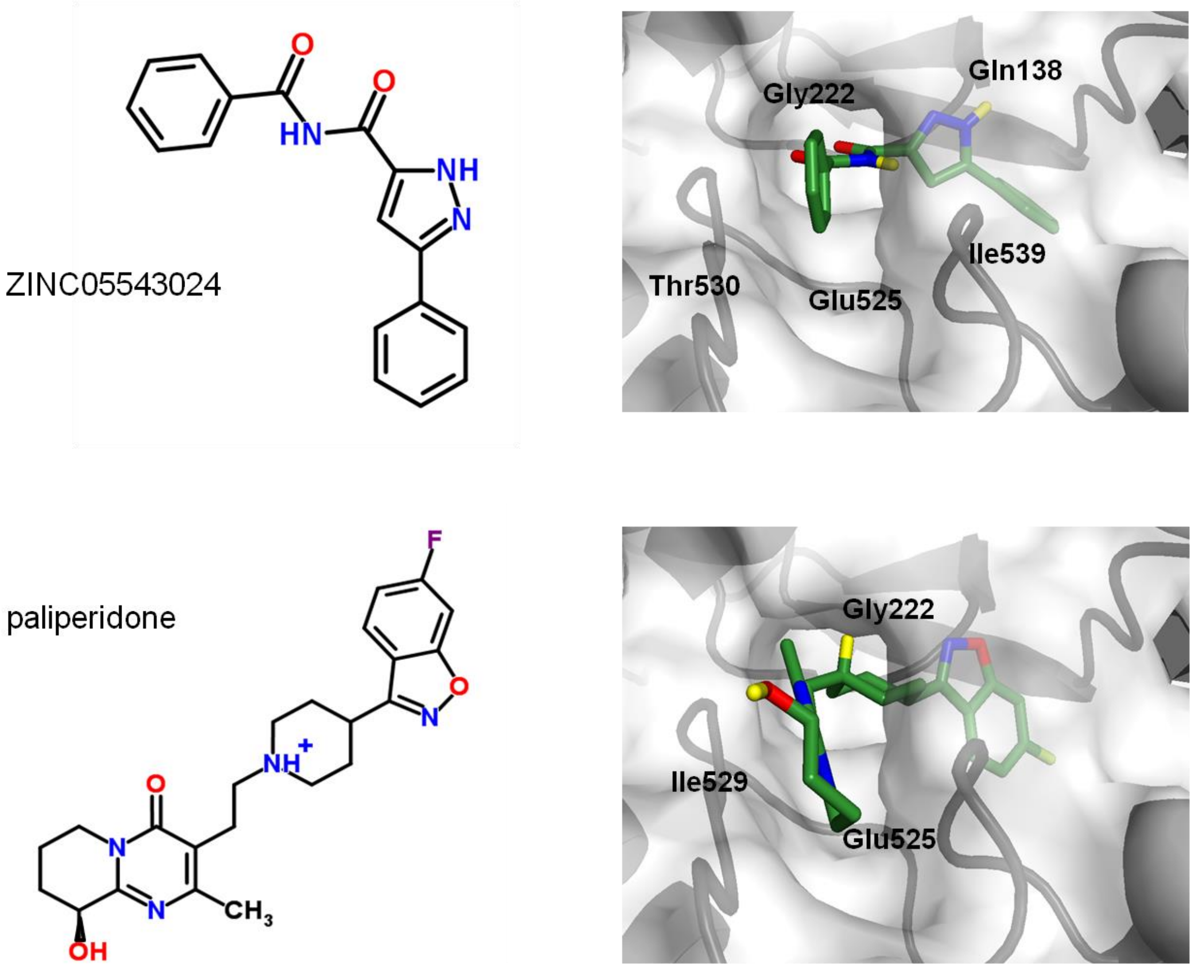


Figure 38. Chemical structure and docking models of top hit compounds targeting the PB2 protein: ZINC05543024 and paliperidone identified by virtual screening using AutoDock Vina. Interacting PB2 residues are labelled. Figure adapted from Patel & Kukol, (2017).

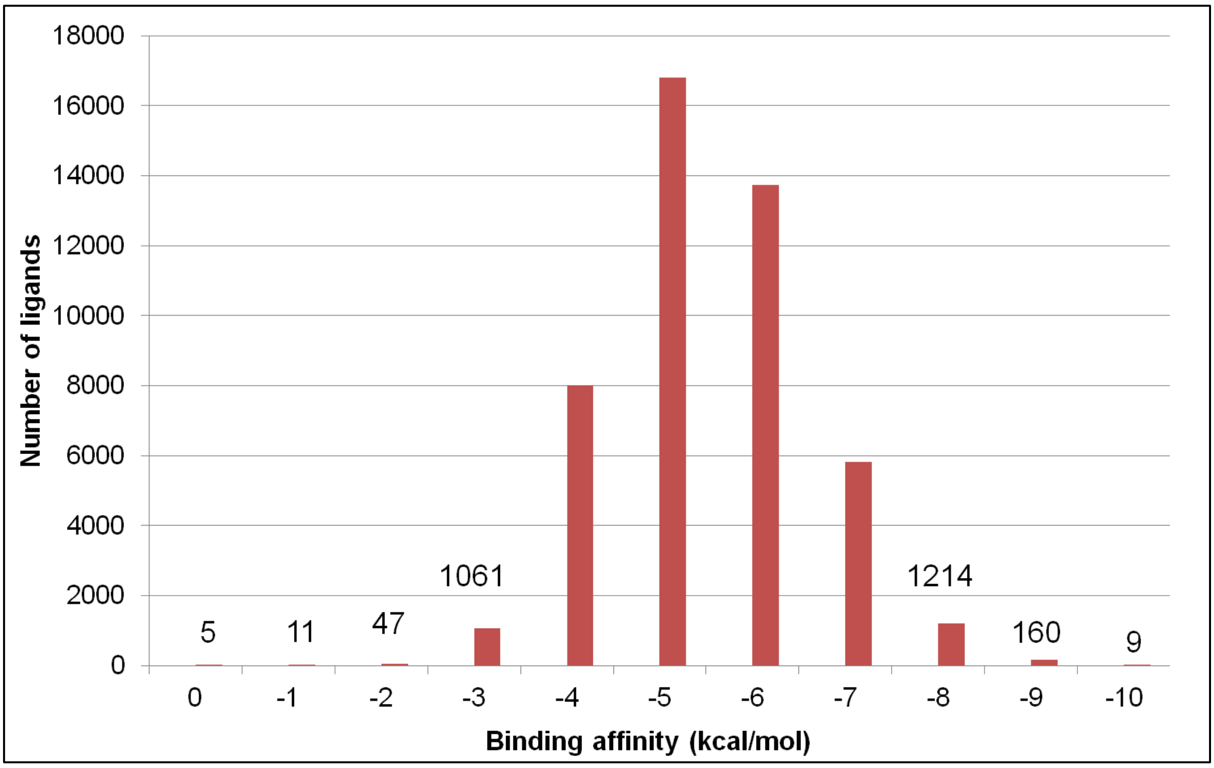


Figure 39. Binding energy distribution graph of compounds screened from the NCI library using AutoDock Vina. (Negative binding scores shown only).

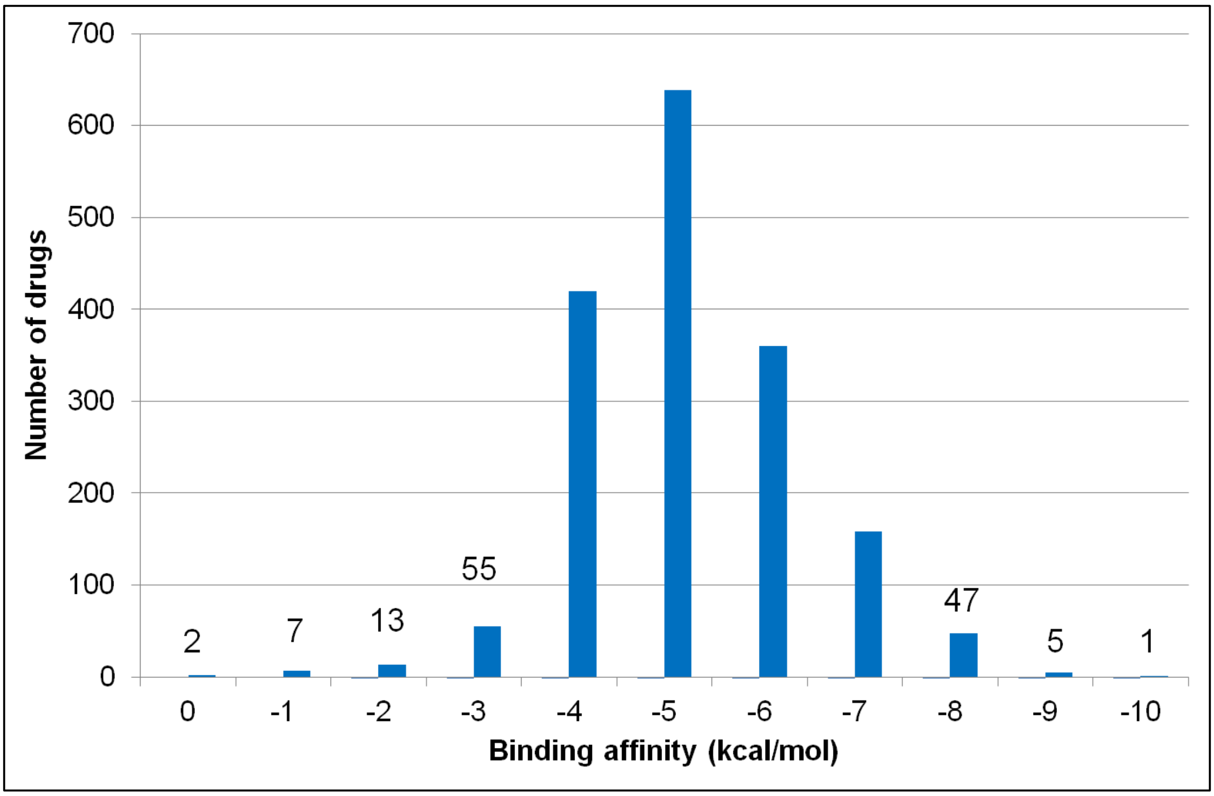


Figure 40. Binding energy distribution graph of compounds screened from the DrugBank-approved library using AutoDock Vina. (Negative scores shown only).

Table 14. Chemical properties and binding affinity (ΔG) of predicted top hit compounds identified from virtual screening of the NCI and DrugBank library obtained from the ZINC database. Properties include molecular mass (Mol M), predicted partition coefficient (xLogP), no. of hydrogen bond donors and acceptors and total polar surface area (tPSA) at pH7.

Compound (ZINC ID)	**ΔG (kcal/mol)**	**Mol M (g/mol)**	**xLogP**	**H Don**	**H Acc**	**tPSA (Å^2)**
NCI library						
ZINC01617371	-10.3	390.42	2.69	1	7	115
ZINC05543024	-10.2	291.31	2.46	2	3	74
ZINC01612458	-10.1	328.76	4.26	2	6	80
ZINC03954617	-10.1	288.31	1.31	2	6	83
ZINC01040450	-10.0	354.32	3.46	2	9	103
ZINC08651894	-10.0	446.93	3.18	2	9	131
ZINC13212434	-10.0	359.84	2.51	4	6	82
ZINC01624487	-9.9	369.83	3.80	1	5	82
ZINC01612446	-9.8	404.73	3.66	2	12	171
ZINC01614027	-9.8	318.40	4.27	1	3	41
DrugBank library						
Paliperidone	-10.0	427.50	1.97	2	7	86
Paliperidone	-9.9	427.50	1.97	2	7	85
Risperidone	-9.4	411.50	2.96	1	6	65
Sulfasalazine	-9.4	397.39	4.08	2	9	147
Folic acid	-9.4	439.39	-2.37	5	13	219
Nebivolol	-9.0	406.45	3.10	4	5	75
Alimta	-8.9	425.40	-1.53	5	11	197
Iloperidone	-8.8	427.50	3.95	1	6	66
Rivaroxaban	-8.8	435.89	2.53	1	8	88
Dantrolene sodium	-8.7	314.26	1.75	1	9	121

5.4 DISCUSSION

5.4.1 PB2 conservation

Regions of structural and functional importance that displayed high conservation include the N-terminal residues 1-37 which form three short α-helices comprising the PB1 binding interface required for effective polymerase activity (Sugiyama et al., 2009). The mGTP cap binding region is also well conserved, albeit with moderately conserved residues at position 339, 340, 453 and 456. Substitution of Lys339 to Thr339 in the cap binding domain of certain subtypes has been found to prevent binding of the phosphate group of mGTP capped mRNA, reducing RNA synthesis, and thereby regulating PB2 activity (Liu et al., 2013). Within this domain, Val414, Arg415 and Gly416 are highly conserved and are required for PB2-acetyl-CoA interaction to maintain transcription activity (Hatakeyama et al., 2014). The 424-loop region is suggested to have an allosteric role in regulating PB1 activity, whilst other conserved residues are expected to contribute to the domains structurally distinct fold which allows formation of intermolecular contacts specific for mGTP cap binding activity (Guilligay et al., 2008).

The 1-269 and 580-683 segments which are reported to be capable of binding the nucleoprotein (NP) (Poole, Elton, Medcalf, & Digard, 2004), also consist of long stretches of conserved residues such as Ser592-Thr612. A total of 42 amino acids were found to be 100% conserved and could therefore be the most resistant to change due to evolutionary adaption of the virus. This includes Leu744 located on a surface exposed loop region, and Gly693, which are suggested to be key residues in the NLS region due to their high conservation, enabling PB2 nuclear entry from the cytoplasm via binding importin-α. Other highly conserved regions with unassigned functions identified in this work may be of interest with regards to further investigations of structure and function of the PB2 protein.

Amino acid residues displaying low or moderate conservation were mostly located on the exterior surface of the protein (figure 35), which is consistent with the finding that surface residues evolve faster than those in the protein core (Warren et al., 2013). The residues neighbouring less conserved positions were generally found to be highly conserved, as well as 16% of residues located in the interior of the protein.

Their restricted variability is presumably essential for maintaining the protein structure, in particular the subdomains.

Due to the majority of sequences being from avian hosts, glutamic acid was the prevalent residue at position 627 based on the consensus sequence. The E627K mutation is well known for determining virulence by increasing polymerase activity and replication in mammals. This prime example of host adaptation is thought to be due to glutamic acid being able to bind the avian version of the host cell factor ANP32A; whereas substitution to lysine allows the polymerase to bind to the mammalian version of this host factor (Long et al., 2016; Moncorgé, Mura, & Barclay, 2010). However, some avian viruses carrying the E627 variant can efficiently replicate in mammalian cells due to compensatory mutations found in the PB1 protein of H5N1 strains (Xu et al., 2012). A mutation study of the 627 domain has also identified specific conserved residues to be essential for general PB2 activity (Arg597, Pro620, Phe621, Arg646 and Arg650), as well as non-essential residues such as Pro625, Pro626 and Gln628, which are also highly conserved (Kirui, Bucci, Poole, & Mehle, 2014). Furthermore, the positive charge of the highly conserved Arg630 (in the presence of nucleoprotein (NP) R150), or Lys627 promotes PB2-NP interaction, which is essential for the ribonucleoprotein complex to provide structural maintenance and regulate viral transcription (Ng et al., 2012). Adaptive mutations to Ala271, Arg591, and Ser590 have also been found to enhance polymerase activity and virus replication in mammals (Bussey, Bousse, Desmet, Kim, & Takimoto, 2010; Mehle & Doudna, 2009; Yamada et al., 2010). The remaining non-conserved positions may also be associated with determining host range, virulence, PB2 cellular localization, or with no particular function.

The protein sequence dataset analysed contains two sequences isolated from bats (including the H17N10 strain for which a crystal structure has been resolved, PDB ID: 4WSB), which are noticeably different to the consensus sequence. Influenza protein sequences isolated from bats have shown less similarity overall to sequences from other hosts (Tong et al., 2013). Despite these differences, the H17N10 sequence for PB2 remains evolutionary close to human and avian strains (Pflug et al., 2014) and is therefore unlikely to result in major structural differences.

5.4.2 Predicted binding hot spots

Computational solvent mapping identified fifteen potential binding hot spots which represent favourable binding regions with small organic molecules. Most of these were located within residues towards the C-terminal part of the polypeptide chain. Based on previous structural information reported and the flexibility of viral polymerase subunits in general (Reich et al., 2014; Thierry et al., 2016), the surface accessibility of some of these binding hot spots may change upon trimer formation or subdomain rotation. Structural alignment of the H5N1 PB2 structure with 4WSB chain C (figure 37b), suggests that the accessibility of these sites would be unaffected as they are not directly blocked by PA/PB1 in complex. Whereas, alternative configurations of the heterotrimer (reviewed by (Pflug, Lukarska, Resa-Infante, Reich, & Cusack, 2017)) have shown that depending on the RNA promoter bound, the PB2 cap-binding, 627 and NLS domains may exist in several states in influenza B and C polymerases. This suggests that accessibility of all hot spots (except for six, nine, ten and thirteen which are not located within these domains) could change.

The highest number of different probes (ten) were found to bind at hot spot one, which indicates that this area has good binding potential with a variety of functional groups. Spots seven, fourteen and fifteen are clustered closely together forming the conserved mGTP cap binding site, and considering the functional importance of this region, the low ranking assigned is probably due to the affinity for the highly charged RNA molecules, which are not well represented by the library of organic solvents used in the docking with the FTmap algorithm. However, these spots are near the binding site for the PB2 inhibitors identified by Clark et al., (2014) (methylguanine derivatives) and Pautus et al., (2013) and consist of residues involved in hydrogen bonding.

Amino acids surrounding spots one, seven, fourteen and fifteen are not involved in heterotrimer formation and are located at positions set apart from the PA/PB1 subunit interactions. Spot six is the only one located within the N-terminal third and could be implicated in trimer association. The conserved region encompassing hot spot two was selected as the target site for docking, as the residues closely surrounding this hot spot display high conservation so are less likely to mutate. Also, this hot spot is second ranked, as several different probe types were predicted to bind there, suggesting it is an important site of the protein (Brenke et al., 2009), and this site has not previously been targeted by virtual screening experiments. Furthermore, there are currently no identified inhibitors which target this hot spot. In relation to PB2 structure and function, the residues surrounding this spot may be associated with rotation of the C-terminal domain, or contribute towards interactions with PB1.

5.4.3 Virtual screening

Some of the key amino acid residues found to interact with the top compounds via hydrogen bonding and hydrophobic interactions using the LigPlot+ software include: Gln138, Gly222, Ile529, Ile539, Asn540, Gly541, Tyr531 and Thr530; all of which are highly conserved. The predicted binding conformations show that some of these compounds bind partially inside a deep pocket formed by these residues between two loop regions (figure 38). Compound 1 (ZINC01617371) forms hydrophobic contacts with eighteen residues and a single hydrogen bond with Ile529 at a distance of 3.12 Å. The compound bends around the 531-541 loop region causing the phenyl ring and nitrile group to be entirely buried within the protein; the methyl group at the other end is surface exposed. Also, three aromatic groups of compound 4 (ZINC03954617) form hydrogen bonds with Gly222, Gln241 and Ile529, and are surrounded by eleven residues forming hydrophobic contacts. The top ten compounds share the common scaffold of an aromatic group at one or both ends, occupying the binding pocket in a similar orientation as compound 2 (ZINC05543024 (figure 38)), supported by van der Waals and electrostatic interactions between atoms. The compounds may inhibit virus replication by interfering with host protein interactions, with trimer assembly by restricting or inducing conformation changes, or with the synthesis of RNA (Thierry et al., 2016); the ligands predicted in this study could serve as tools to investigate such functions, or be used as a starting point to develop clinically active anti-influenza drugs.

The drug paliperidone is approved by the Food and Drug Administration (FDA) for the treatment of schizophrenia and related disorders. It binds to the dopamine and serotonin receptors, although the exact mechanism of action is not known (reviewed in Corena-McLeod, (2015)). Paliperidone is a large compound occupying most of the binding pocket; the nitrogen atom of the central pyridine ring is able to form a hydrogen bond with the oxygen atom of Glu241, and the drug-protein complex is maintained via hydrophobic contacts with sixteen surrounding residues of the target site. The results of this study may be useful for repurposing this drug or derivatives as a treatment for influenza A infection.

6 EVALUATION OF A NOVEL VIRTUAL SCREENING STRATEGY USING RECEPTOR DECOY BINDING SITES

6.1 INTRODUCTION

This chapter describes the development and evaluation of a novel yet straightforward receptor-decoy strategy that aimed to improve molecular docking predictions (Patel & Kukol, 2016a). Several tools for molecular docking and virtual screening are available (Plewczynski, Łaźniewski, Augustyniak, & Ginalski, 2011) and have been evaluated on various protein-ligand complexes (Li, Li, Cheng, Liu, & Wang, 2010; Tuccinardi, Poli, Romboli, Giordano, & Martinelli, 2014). Although docking provides an efficient and cost effective way to assess interactions between molecules such as proteins and ligands on a large-scale, the accuracy, as defined by the ability to predict strong binding ligands, is limited. This is largely due to the limitation of scoring functions used in the software to calculate binding energies, and therefore their ability to identify true positives from a database composed of known ligands and decoys (molecules with physical properties similar to known ligands but dissimilar topology) that is typically used in evaluations of virtual screening (Huang, Shoichet, & Irwin, 2006; Kitchen et al., 2004). The accuracy of the screening method can be assessed quantitatively through calculation of the robust metric known as Receiver Operator Characteristic Enrichment (ROCE) (Nicholls, 2008). An ROCE factor is obtained as the true positive rate divided by the false positive rate, therefore ROCE factors much larger than 1.0 are desirable to establish that the docking algorithm can distinguish active compounds from decoys.

Methods to increase the accuracy of virtual screening have previously been suggested, for example, considering receptor flexibility to reduce the numbers of false positive molecules (Awuni & Mu, 2015), consensus docking to predict correct binding pose (Houston & Walkinshaw, 2013), and a consensus virtual screening method that combined the rank lists of ligands from up to three different algorithms (Kukol, 2011). However, these improved methods can still result in a low number of correct predictions for some receptors.

6.1.1 Aims and objectives

This work aimed to improve the accuracy of molecular docking predictions, thus supporting the discovery of influenza A virus replication inhibitors. This was attempted by removing possible false positive compounds, which involves performing virtual screening against a non-binding (decoy) site on the same receptor protein target. A method to re-rank the screening results was developed, enabling a comparison of ROCE factors before and after the application of receptor-decoy screening in order to evaluate the novel strategy.

6.2 MATERIALS AND METHODS

6.2.1 Materials

Ligand and decoy sets for fifteen target proteins were downloaded from the Database of Useful Decoys (http://dud.docking.org) (Huang et al., 2006). The complexes were selected from several different protein categories in the database such as hormone receptors, kinases, proteases and other enzymes to represent a wide range of targets, including ten targets which had previously been used to evaluate common docking algorithms (Kukol, 2011). Ligand and decoy sets were converted from the .mol2 format to the PDBQT format using the screening preparation software Raccoon. The FTMap server (Brenke et al., 2009) was used to define the decoy site for docking. Virtual screening for all fifteen targets was performed using AutoDock Vina version 1.1.1 with the default parameters on the University of Hertfordshire high performance computer cluster.

6.2.2 Docking against the binding site and decoy site

For each target, a text file with a list of decoys and a list of ligands was generated. The search space for docking was defined via a grid box manually specified with AutoDock Tools (Morris & Huey, 2009) around the binding or decoy site. A grid spacing of 0.375 Å was used to determine the box dimensions which remained the same for the binding and decoy site. The decoy site was chosen based on the following criteria: 1) contains no binding hot spot predicted by FTMap, 2) it appears structurally different to the actual binding site and 3) it does not form an obvious binding cavity, but is at a flat region on the exterior surface of the protein.

6.2.3 Generating adjusted rank lists

A script was used to generate adjusted rank lists from the binding site list by considering molecules that were in the top 10%, 15%, 20%, 30% and 50% of the decoy site list, and adjusting the rank of the binding site list based on the following formula:

$$New\ rank = (Binding\ site\ rank - Decoy\ site\ rank) + \ Total\ no.of\ ligands\ in\ list$$

As a result, molecules that received a high rank in the screening against the decoy site were assigned a much lower rank in the screening results against the binding site.

The fraction of decoy-site docking results was varied in order to find a cut-off where maximum enrichment is achieved. With the example of the progesterone receptor (PR), the following command was used taking into account the top 10% of the decoy-site rank list with a weight of 1.0:

```
$ perl RankAdjustDecoySite.pl pr_bindingsitelist.txt
pr_decoysitelist.txt 10 1.0 > PR_decoyAdjust_10.txt
```

The weight was always left at 1.0.

6.2.4 Calculation of Receiver Operator Characteristic Enrichment (ROCE)

The numbers of active ligands in the database were then used to calculate the ROCE factors at 1% and 2% of the number of molecules in each of the adjusted rank lists using a script. With the example of the receptor PR, the following command was used:

```
$ perl CountLigands.pl PR_ligandsList.txt PR_decoyAdjust_10.txt
(top)
```

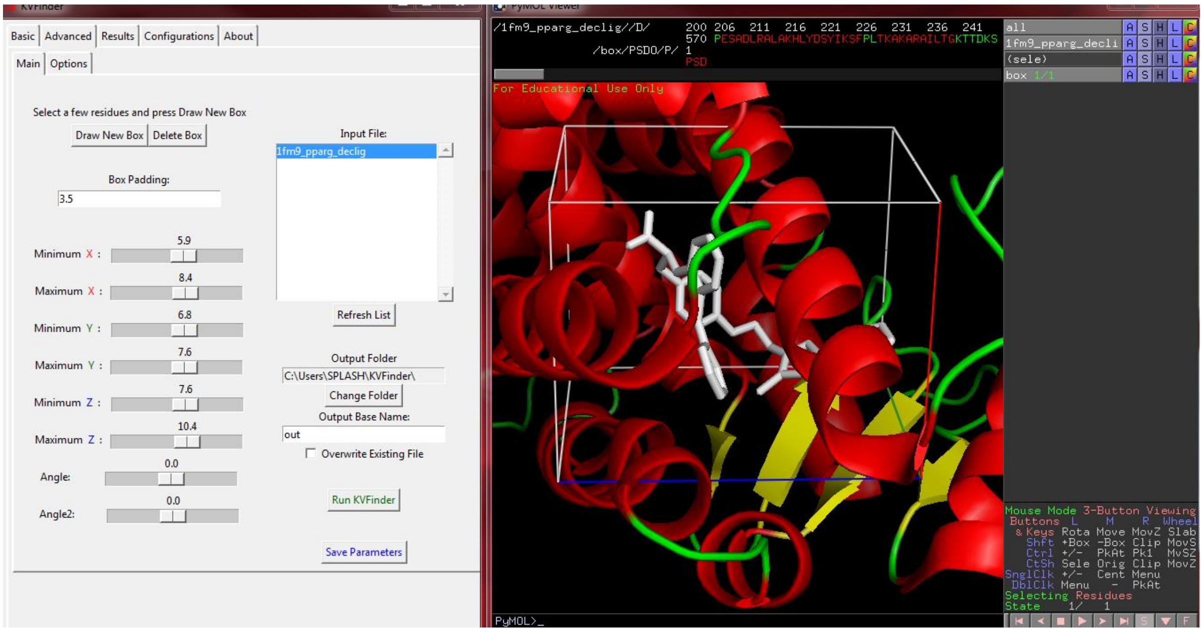
The ROCE was calculated as the fraction of true positives divided by the fraction of false positives at x% of the ligand/decoy database according to the equation:

$$ROCE_{x\%} = \frac{f_{actives}}{1 - \frac{(N_{decoys} - N_{inactives})}{N_{decoys}}}$$

Where $f_{actives}$ = (number of actives at x%) / (number of all actives), N_{decoys} = the total number of inactive decoys, $N_{inactives}$ = the number of decoys chosen at x% of the ligand/decoy database.

6.2.5 Binding site and decoy site analysis

Binding site and decoy sites were analysed post-docking with the KVFinder Cavity Detection PyMol Plugin (Oliveira et al., 2014) to provide a quantitative description of the two sites. The software enables comparison and characterisation of protein binding sites by the number, area and volume of cavities in a specified search space. The default parameters were used for all fifteen targets, which included a probe in size of 1.4 Å, probe out size of 4.0 Å and a step size of 0.6 Å. The minimum cavity volume was set at 5.0 Å. The binding site search space was set around the position of the actual ligand molecule obtained from the Protein Data Bank (as shown in figure 41), and the decoy site search space was set using a docked molecule from the decoy site screening.



Select a few residues and press Draw New Box

Box Padding: 3.5

Minimum X : 5.9

Maximum X : 8.4

Minimum Y : 6.8

Maximum Y : 7.6

Minimum Z : 7.6

Maximum Z : 10.4

Angle: 0.0

Angle2: 0.0

Input File: 1fm9_pparg_declig

Output Folder: C:\Users\SPLASH\KVFinder\

Output Base Name: out

□ Overwrite Existing File

```
/1fm9_pparg_declig//D/     200 206  211  216  221  226  231  236  241
                           570 PESADLRALAKHLYDSYIKSFPLTKAKARAILTGKTTDKS
              /box/PSDO/P/ 1
                           PSD
For Educational Use Only
PyMOL>_
```

Figure 41. Screenshot of the KVFinder cavity detection software used to analyse the binding site for the protein Pparg. The box dimensions were set around the position of the actual ligand.

6.3 RESULTS

6.3.1 ROCE at different fractions of the adjusted binding site rank list

In this study, the level of Receiver Operator Characteristic Enrichment (ROCE) was determined at fractions of 1% and 2% of the dataset of ligand/decoy molecules. Performing the docking against a decoy site on the same receptor, as shown in figure 42, lead to a ranking of molecules different from the ranking for the true binding site. The predicted binding energies among top molecules for the decoy site were less negative than for binding sites, indicating a lower degree of binding to the decoy site. The ranking for the true binding site was adjusted by considering a varied fraction of the rank list produced from the decoy site from 0% (no correction) to 50% (table 15 and 16). Overall, the majority of targets did not show any improvement in enrichment at the top 1% or 2% of the list after applying the receptor decoy method. At 1% of the database, five targets (Comt, Ache, CDK2, HIVrt and Pparg) show improved ROCE factors compared to those obtained in the previous study (Kukol, 2011 (see footnotes in table 15 and 16)), when considering at least the top 15% of the decoy site list. Beyond 15% the enrichment for most targets either remained constant or dropped to a lower value.

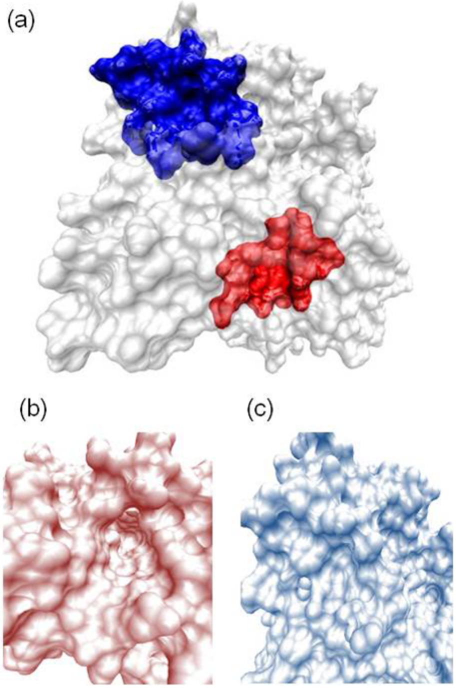


Figure 42. (a) Ache receptor with the true binding site shown in red and decoy site in blue. (b) detailed view of the Ache binding site. (c) Detailed view of the Ache decoy site. Figure reproduced from Patel & Kukol (2016a).

Table 15. ROCE at 1% of the binding site list considering top x% of the decoy site list. Numbers in bold indicate improvement over the unadjusted virtual screening.

	Top % of decoy list					
Receptor	**0**	**10**	**15**	**20**	**30**	**50**
Comt	33.4[1]	39.0	**39.0**	39.0	39.0	39.0
AchE	1.0[1]	0.96	**3.03**	3.03	1.97	3.03
CDK2	14.3[1]	23.6	**29.7**	23.6	29.7	14.3
HIVrt	9.4[1]	13.1	**13.1**	9.0	13.1	18.0
Pparg	58.8[1]	99.0	**84.0**	84.0	84.0	84.0
FGFR1	0.0	0.0	0.9	0.9	0.9	0.9
InhA	14.6	5.3	5.3	5.3	2.5	0.0
PR	0.0	0.0	0.0	0.0	0.0	4.0
RXRa	212.4	212.4	212.4	88.5	47.2	88.5
VEGFr2	10.2	2.9	2.9	2.9	1.4	1.4
MR	178.3	178.3	71.3	71.3	71.3	7.1
Hsp90	0.0	0.0	0.0	0.0	0.0	0.0
ampC	0.0	0.0	0.0	0.0	0.0	0.0
trypsin	5.0	5.0	5.0	5.0	5.0	5.0
Parp	7.1	7.1	7.1	7.1	7.1	11.9

[1] Results taken from (Kukol, 2011)

Table 16. ROCE at 2% of the binding site list considering top x% of the decoy site list. Numbers in bold indicate improvement over the unadjusted virtual screening.

	Top % of decoy list					
Receptor	**0**	**10**	**15**	**20**	**30**	**50**
Comt	10.4[1]	**11.2**	11.2	11.2	11.2	11.2
AchE	1.5[1]	**2.0**	2.0	1.5	2.0	1.5
CDK2	13.3[1]	**15.7**	15.7	15.7	11.9	5.8
HIVrt	8.8[1]	**9.0**	9.0	7.2	7.2	11.0
Pparg	35.8[1]	**58.0**	54.0	58.0	50.0	44.0
FGFR1	0.6	0.4	0.4	0.4	0.4	0.4
InhA	12.4	3.1	3.1	2.5	1.2	0.0
PR	0.0	0.0	0.0	0.0	1.9	4.0
RXRa	70.8	**97.4**	70.8	40.5	17.7	23.6
VEGFr2	5.3	1.4	2.1	2.1	2.9	1.4
MR	62.4	42.8	29.7	20.4	20.4	3.6
Hsp90	0.0	0.0	0.0	0.0	0.0	4.5
ampC	0.0	0.0	0.0	0.0	0.0	0.0
trypsin	5.0	5.0	5.0	5.0	5.0	5.0
Parp	5.1	3.2	7.1	7.1	7.1	9.4

[1] Results taken from (Kukol, 2011)

The largest improvement in enrichment was achieved with the targets CDK2 and Pparg. Figure 43 shows that the ROC curve with decoy site adjustment is either similar, or even below the unadjusted curve for the target CDK2. However, it is the early enrichment that is important when utilising virtual screening results for experimental testing (inset to figure 43). For the targets PR and Hsp90, the ROCE at 1% and 2% remained at zero until considering at least 30% of molecules in the decoy list.

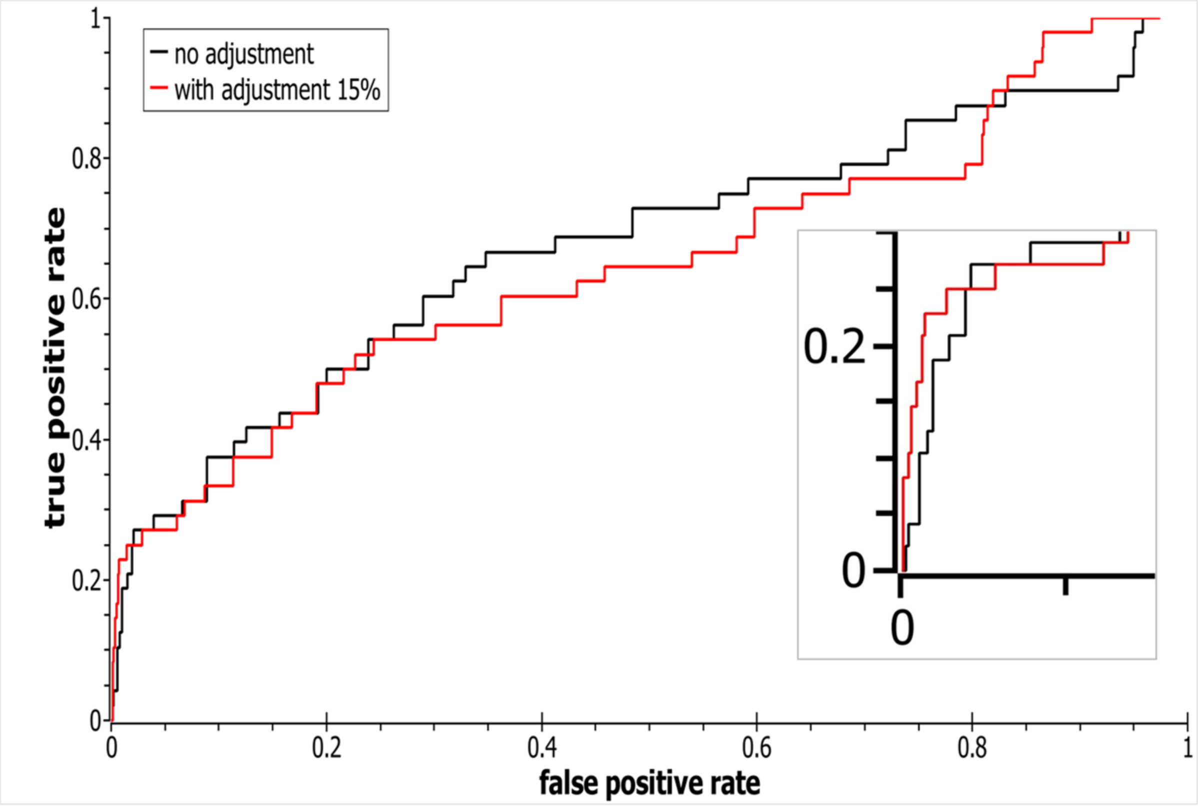


Figure 43. Receiver Operator Characteristic (ROC) curve for the target CDK2 showing improved early enrichment with decoy site adjustment of 15% compared to no adjustment. The inset shows the early enrichment up to 5% of ligand/decoy molecules.

6.3.2 Binding site and decoy site analysis

Cavity analyses of the binding site and decoy site (table 17) using the software KVFinder (Oliveira et al., 2014) shows that the total number, volume and area of the cavities found in the decoy site were smaller in comparison to the binding site for all targets, except HIVrt and trypsin. This confirms that the shapes of the two sites are very different; although this does not prevent false positive molecules binding with high affinity.

Table 17. Analysis of binding and decoy sites for all receptors based on the number, volume and area of cavities in the search space using KVFinder.

	Binding site Cavities			**Decoy site cavities**		
Receptor	**Number**	**Total Volume (Å^3)**	**Total Area (Å^2)**	**Number**	**Total Volume (Å^3)**	**Total Area (Å^2)**
Comt	1	29.8	45.7	1	12.3	20.3
AchE	3	249.3	333.8	2	85.5	124.9
CDK2	4	134.0	178.3	1	10.6	15.8
HIVrt	5	92.1	138.7	1	241.3	240.1
Pparg	2	394.4	414.8	1	8.6	14.4
FGFR1	2	49.0	70.2	1	21.0	30.6
InhA	2	1119.7	834.8	1	6.0	10.1
PR	2	21.4	35.0	1	18.1	28.4
RXRa	1	57.5	72.0	1	21.0	30.2
Vegfr2	5	129.3	193.0	4	117.4	168.0
MR	2	54.4	78.1	1	23.5	35.6
Hsp90	4	166.8	233.7	2	30.0	46.1
ampC	3	100.8	121.3	1	5.8	9.7
trypsin	1	9.7	14.8	3	79.9	121.8
Parp	1	538.3	482.4	1	9.9	16.6

6.4 DISCUSSION

High predicted binding affinities between a ligand and a receptor may not always correspond with the best binding molecules for the target site investigated (Plewczynski et al., 2011; Wang, Lu, Fang, & Wang, 2004). In virtual screening, this is reflected by low enrichment factors, which indicate that many of the highest ranked molecules may be false positive predictions (Nicholls, 2008). The rationale behind the receptor decoy strategy was that the number of false positive binders could be reduced by determining molecules, which have a tendency to bind non-specifically to molecular surfaces. As a result, a higher number of active ligands would remain after processing the rank list for the true binding site with the rank list for the decoy site. However, the results show that this approach is unlikely to help in the identification and selection of molecules for experimental testing as a higher number of true positives were recalled for only five targets.

The extent of enrichment achieved for the top 1% and 2% differed for all targets due to properties that determine the binding interactions between amino acid residues of the target and the ligand-decoy dataset used for docking. The optimum cut-off for maximum enrichment at the top 1% of a binding site list was obtained when considering 15% of the decoy list (table 15), and 10% for the top 2% of the binding site list (table 16). This shows that the ranking of molecules with regards to binding to the decoy sites is meaningless for lower ranks. For those targets where the ROCE factor remained at zero until considering at least 30% of the decoy list, indicates that true and false ligands cannot be distinguished by the AutoDock Vina docking algorithm.

The results show a considerable variation between the fifteen targets investigated confirming the general consensus that virtual screening accuracy is highly dependent on the target. The targets Inha, MR and VEGFr2 show a significant decrease in ROCE, indicating this strategy makes the retrieval of active ligands in the top ranks worse for these targets. The actual binding site for VEGFr2 appears to be non-specific, open and flat, therefore binds molecules which also bind easily to the decoy site, resulting in a high proportion of active molecules at the top of the decoy list. However, the Inha binding site is a small, deep pocket with a total cavity area of 838.4 Å^2 which appears not to be easily surface accessible, so it is expected that this receptor only binds ligands which are complementary in shape. Although, this was not seen as a higher number of active ligands were found in the top 1% of the decoy site list compared to the binding site list. Thus, when the re-ranking formula to generate the adjusted list is applied, the binding site list is re-ordered such that the active ligands do not appear in the top positions. This highlights the shortcoming that if applying this strategy to a virtual screening experiment where active molecules are not known, it cannot be guaranteed that any improved prediction accuracy will result.

6.4.1 Conclusion

The novel development and evaluation of docking with a decoy binding site shows that improved prediction of active ligands could not be achieved in general. It should be noted that the ligand/decoy dataset used for this evaluation is especially challenging as decoys with physico-chemical properties similar to ligands were chosen (Huang et al., 2006). The choice of appropriate decoy binding sites is critical for the success of this method. Choosing an obviously unfavorable site, such as a flat molecular surface, reduces the docking scores overall, and thus the potential to discriminate between ligands and decoys, while on the other hand, the choice of an alternative binding cavity might cause a novel mode of specific binding that does help to eliminate the false positives for the true binding site. The question, how to define a decoy binding site, such that false positive predictions for the real binding site are removed must remain. Further work addressing the re-ranking of predicted ligands may also lead to improvements.

7 OVERALL CONCLUSION

The influenza A viruses are capable of undergoing rapid evolution due to the high error rate during the replication cycle, as well as the ability of undergoing genetic re-assortments. This not only allows the virus to evade selective pressures such as the host immune response, but also susceptibility to antiviral drugs. Therefore this work has aimed to identify drug like compounds which may bind to the most evolutionary stable binding sites on the NS1, NEP and PB2 internal proteins.

Taken together, the results reveal which specific regions of the NS1, NEP and PB2 proteins display the highest levels of sequence conservation and how these regions are implicated in protein function and the virus life cycle overall. Furthermore, the analysis of predicted binding hot spots in this work has focussed on targeting sites other than those frequently reported in literature, such as the NS1-CPSF30 and PB2-mGTP RNA cap snatching interaction sites (Byrn et al., 2015; Engel, 2013; Pautus et al., 2013; Twu et al., 2006). Starting with the NS1 protein as a potential target, two out of twelve predicted top hit compounds showed inhibition of H1N1 virus replication through plaque reduction, although their predicted target protein could not be confirmed. Also, new structural insights were revealed for the NEP, which showed that despite being a stable structure, subtle changes in side chain orientations can influence ligand binding hot spot locations. Overall, the predicted binding affinities from the docking experiments with compounds from the NCI library were higher for the PB2 protein (up to -10.3 kcal/mol) compared to the NEP, which suggests that the PB2 hot spot targeted could be a more favorable drug target site.

Limitations in this work include the accuracy of the NS1 and NEP models and their conformations sampled through MD simulations, as well as the accuracy of docking predictions, as discussed in chapter six. Another limiting factor relating to the discovery of inhibitors is the compatibility of the top compounds identified with the antiviral assay used for validation, as shown in chapter three. The specificity towards the target protein is also important as the identified compounds may bind and block other viral proteins displaying similar structure and properties to the target site investigated.

In conclusion, this research has identified drug-like compounds predicted to bind with strong affinity to a region of the NS1, NEP and PB2 proteins consisting of highly conserved amino acid residues. Due to the low probability of the targeted regions undergoing genetic changes amongst different virus subtypes and hosts, such compounds may remain viable long-term as universal influenza inhibitors.

Future work should be directed towards *in vitro* investigations to verify inhibition on virus replication and whether the predicted compounds could be developed into successful antiviral drugs. Further experiments are required to elucidate the antiviral mechanism of action for compounds D and K. Also, studies to test for the development of resistance by serial passage of the virus in cell culture with increasing drug compound concentrations could be performed to establish that influenza A viruses do not become resistant against the compounds identified. Potential highly conserved binding sites which are yet to be characterised could be investigated in research aimed at a further understanding of influenza virus biology and virus-host interactions.

8 REFERENCES

Akarsu, H., Burmeister, W. P., Petosa, C., Petit, I., Müller, C. W., Ruigrok, R. W. H., et al. (2003). Crystal structure of the M1 protein-binding domain of the influenza A virus nuclear export protein (NEP/NS2). *The EMBO Journal*, *22*(18), 4646–55. https://doi.org/10.1093/emboj/cdg449

Amorim, M. J., Kao, R. Y., & Digard, P. (2013). Nucleozin targets cytoplasmic trafficking of viral ribonucleoprotein-Rab11 complexes in influenza A virus infection. *Journal of Virology*, *87*(8), 4694–703. https://doi.org/10.1128/JVI.03123-12

Awuni, Y., & Mu, Y. (2015). Reduction of false positives in structure-based virtual screening when receptor plasticity is considered. *Molecules (Basel, Switzerland)*, *20*(3), 5152–64. https://doi.org/10.3390/molecules20035152

Ayllon, J., Russell, R. J., García-Sastre, A., & Hale, B. G. (2012). Contribution of NS1 effector domain dimerization to influenza A virus replication and virulence. *Journal of Virology*, *86*(23), 13095–8. https://doi.org/10.1128/JVI.02237-12

Baell, J. B., & Holloway, G. A. (2010). New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *Journal of Medicinal Chemistry*, *53*(7), 2719–2740. https://doi.org/10.1021/jm901137j

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., et al. (2008). The influenza virus resource at the National Center for Biotechnology Information. *Journal of Virology*, *82*(2), 596–601. https://doi.org/10.1128/JVI.02005-07

Basu, D., Walkiewicz, M. P., Frieman, M., Baric, R. S., Auble, D. T., & Engel, D. a. (2009). Novel influenza virus NS1 antagonists block replication and restore innate immune function. *Journal of Virology*, *83*(4), 1881–1891. https://doi.org/10.1128/JVI.01805-08

Beauchamp, K. A., Lin, Y. S., Das, R., & Pande, V. S. (2012). Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *Journal of Chemical Theory and Computation*, *8*(4), 1409–1414.

Bernardi, R. C., Melo, M. C. R., & Schulten, K. (2015). Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta*, *1850*(5), 872–7. https://doi.org/10.1016/j.bbagen.2014.10.019

Bornholdt, Z. A., & Prasad, B. V. V. (2008). X-ray structure of NS1 from a highly pathogenic H5N1 influenza virus. *Nature*, *456*(7224), 985–8. https://doi.org/10.1038/nature07444

Bouvier, N. M., & Palese, P. (2008). THE BIOLOGY OF INFLUENZA VIRUSES. *Vaccine*, *26(Suppl 4*, D49-53.

Boyd, M. J., Bandarage, U. K., Bennett, H., Byrn, R. R., Davies, I., Gu, W., et al. (2015). Isosteric replacements of the carboxylic acid of drug candidate VX-787: Effect of charge on antiviral potency and kinase activity of azaindole-based influenza PB2 inhibitors. *Bioorganic & Medicinal Chemistry Letters*, *25*(9), 1990–4. https://doi.org/10.1016/j.bmcl.2015.03.013

Brenke, R., Kozakov, D., Chuang, G. Y., Beglov, D., Hall, D., Landon, M. R., et al. (2009). Fragment-based identification of druggable “hot spots” of proteins using Fourier domain correlation techniques. *Bioinformatics*, *25*(5), 621–627. https://doi.org/10.1093/bioinformatics/btp036

Bussey, K. A., Bousse, T. L., Desmet, E. A., Kim, B., & Takimoto, T. (2010). PB2 residue 271 plays a key role in enhanced polymerase activity of influenza A viruses in mammalian host cells. *Journal of Virology*, *84*(9), 4395–406. https://doi.org/10.1128/JVI.02642-09

Byrn, R. A., Jones, S. M., Bennett, H. B., Bral, C., Clark, M. P., Jacobs, M. D., et al. (2015). Preclinical activity of VX-787, a first-in-class, orally bioavailable inhibitor of the influenza virus polymerase PB2 subunit. *Antimicrobial Agents and Chemotherapy*, *59*(3), 1569–82. https://doi.org/10.1128/AAC.04623-14

Capra, J. a., & Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics*, *23*(15), 1875–1882. https://doi.org/10.1093/bioinformatics/btm270

Carrat, F., & Flahault, A. (2007). Influenza vaccine: The challenge of antigenic drift. *Vaccine*, *25*(39–40), 6852–6862. https://doi.org/10.1016/j.vaccine.2007.07.027

Carrillo, B., Choi, J.-M., Bornholdt, Z. a, Sankaran, B., Rice, A. P., & Prasad, B. V. V. (2014). The influenza A virus protein NS1 displays structural polymorphism. *Journal of Virology*, *88*(8), 4113–22. https://doi.org/10.1128/JVI.03692-13

CDC. (2014). Transmission of Influenza Viruses from Animals to People. Retrieved May 17, 2016, from http://www.cdc.gov/flu/about/viruses/transmission.htm

CDC. (2016). Influenza Antiviral Medications: Summary for Clinicians. Retrieved May 22, 2016, from http://www.cdc.gov/flu/professionals/antivirals/summary-clinicians.htm

CDC. (2017). Seasonal Influenza Vaccine Effectiveness, 2005-2017. Retrieved July 1, 2017, from https://www.cdc.gov/flu/professionals/vaccination/effectiveness-studies.htm

Chambers, B. S., Parkhouse, K., Ross, T. M., Alby, K., & Hensley, S. E. (2015). Identification of Hemagglutinin Residues Responsible for H3N2 Antigenic Drift during the 2014-2015 Influenza Season. *Cell Reports*, *12*(1), 1–6. https://doi.org/10.1016/j.celrep.2015.06.005

Chen, J., Huang, S., & Chen, Z. (2010). Human cellular protein nucleoporin hNup98 interacts with influenza A virus NS2/nuclear export protein and overexpression of its GLFG repeat domain can inhibit virus propagation. *Journal of General Virology*, *91*(10), 2474–2484. https://doi.org/10.1099/vir.0.022681-0

Chen, Y. C. (2015). Beware of docking! *Trends in Pharmacological Sciences*, *36*(2), 78–95. https://doi.org/10.1016/j.tips.2014.12.001

Chen, Z., Li, Y., & Krug, R. M. (1999). Influenza A virus NS1 protein targets poly(A)-binding protein II of the cellular 3'-end processing machinery. *The EMBO Journal*, *18*(8), 2273–83. https://doi.org/10.1093/emboj/18.8.2273

Cheng, A., Wong, S. M., & Yuan, Y. A. (2009). Structural basis for dsRNA recognition by NS1 protein of influenza A virus. *Cell Research*, *19*, 187–195. https://doi.org/10.1038/cr.2008.288

Cheung, T. K. W., & Poon, L. L. M. (2007). Biology of influenza a virus. In *Annals of the New York Academy of Sciences* (Vol. 1102, pp. 1–25).

Cho, E. J., Xia, S., Ma, L.-C., Robertus, J., Krug, R. M., Anslyn, E. V., et al. (2012). Identification of Influenza Virus Inhibitors Targeting NS1A Utilizing Fluorescence Polarization-Based High-Throughput Assay. *Journal of Biomolecular Screening*, *17*, 448–459. https://doi.org/10.1177/1087057111431488

Clark, M. P., Ledeboer, M. W., Davies, I., Byrn, R. A., Jones, S. M., Perola, E., et al. (2014). Discovery of a novel, first-in-class, orally bioavailable azaindole inhibitor (VX-787) of influenza PB2. *Journal of Medicinal Chemistry*, *57*(15), 6668–6678. https://doi.org/10.1021/jm5007275

Corena-McLeod, M. (2015). Comparative Pharmacology of Risperidone and Paliperidone. *Drugs in R&D*, *15*(2), 163–174. https://doi.org/10.1007/s40268-015-0092-x

Couch, R. B. (1996). Orthomyxoviruses. In S. Baron (Ed.), *Medical Microbiology* (4th ed.). University of Texas Medical Branch at Galveston.

Darapaneni, V., Prabhaker, V. K., & Kukol, A. (2009). Large-scale analysis of influenza A virus sequences reveals potential drug target sites of non-structural proteins. *The Journal of General Virology*, *90*, 2124–2133. https://doi.org/10.1099/vir.0.011270-0

Das, K., Aramini, J. M., Ma, L.-C., Krug, R. M., & Arnold, E. (2010). Structures of influenza A proteins and insights into antiviral drug targets. *Nature Structural & Molecular Biology*, *17*(5), 530–538.

Das, K., Ma, L.-C., Xiao, R., Radvansky, B., Aramini, J., Zhao, L., et al. (2008). Structural basis for suppression of a host antiviral response by influenza A virus. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(35), 13093–13098.

Dawood, F. S., Iuliano, A. D., Reed, C., Meltzer, M. I., Shay, D. K., Cheng, P.-Y., et al. (2012). Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. *The Lancet. Infectious Diseases*, *12*(9), 687–95. https://doi.org/10.1016/S1473-3099(12)70121-4

DeDiego, M. L., Nogales, A., Lambert-Emo, K., Martinez-Sobrido, L., & Topham, D. J. (2016). NS1 Protein Mutation I64T Affects Interferon Responses and Virulence of Circulating H3N2 Human Influenza A Viruses. *Journal of Virology*, *90*(21), 9693–9711. https://doi.org/10.1128/JVI.01039-16

Delaforge, E., Milles, S., Bouvignies, G., Bouvier, D., Boivin, S., Salvi, N., et al. (2015). Large-Scale Conformational Dynamics Control H5N1 Influenza Polymerase PB2 Binding to Importin α. *Journal of the American Chemical Society*, *137*(48), 15122–15134. https://doi.org/10.1021/jacs.5b07765

Engel, D. A. (2013). The influenza virus NS1 protein as a therapeutic target. *Antiviral Research*, *99*(3), 409–16. https://doi.org/10.1016/j.antiviral.2013.06.005

Englund, J. A. (2002). Antiviral therapy of influenza. *Seminars in Pediatric Infectious Diseases*, *13*(2), 120–128. https://doi.org/10.1053/spid.2002.122999

Fodor, E. (2013). The RNA polymerase of influenza A virus: mechanisms of viral transcription and replication. *Acta Virologica*, *57*(1), 113–122. https://doi.org/10.4149/av

Forli, S., Huey, R., Pique, M. E., Sanner, M. F., Goodsell, D. S., & Olson, A. J. (2016). Computational protein–ligand docking and virtual drug screening with the AutoDock suite. *Nature Protocols*, *11*(5), 905–919. https://doi.org/10.1038/nprot.2016.051

Furuta, Y., Gowen, B. B., Takahashi, K., Shiraki, K., Smee, D. F., & Barnard, D. L. (2013). Favipiravir (T-705), a novel viral RNA polymerase inhibitor. *Antiviral Research*, *100*(2), 446–54. https://doi.org/10.1016/j.antiviral.2013.09.015

Gack, M. U., Albrecht, R. A., Urano, T., Inn, K.-S., Huang, I.-C., Carnero, E., et al. (2009). Influenza A virus NS1 targets the ubiquitin ligase TRIM25 to evade recognition by the host viral RNA sensor RIG-I. *Cell Host & Microbe*, *5*(5), 439–49. https://doi.org/10.1016/j.chom.2009.04.006

Gao, S., Wang, S., Cao, S., Sun, L., Li, J., Bi, Y., et al. (2014). Characteristics of nucleocytoplasmic transport of H1N1 influenza A virus nuclear export protein. *Journal of Virology*, *88*(13), 7455–63. https://doi.org/10.1128/JVI.00257-14

Gao, S., Wu, J., Liu, R.-Y., Li, J., Song, L., Teng, Y., et al. (2015). Interaction of NS2 with AIMP2 facilitates the switch from ubiquitination to SUMOylation of M1 in influenza A virus-infected cells. *Journal of Virology*, *89*(1), 300–11. https://doi.org/10.1128/JVI.02170-14

Gedeon, P. C., Thomas, J. R., & Madura, J. D. (2015). Accelerated Molecular Dynamics and Protein Conformational Change: A Theoretical and Practical Guide Using a Membrane Embedded Model Neurotransmitter Transporter. In A. Kukol (Ed.), *Molecular Modeling of Proteins* (2nd ed., pp. 253–287). Humana Press. https://doi.org/10.1007/978-1-4939-1465-4_12

Gorai, T., Goto, H., Noda, T., Watanabe, T., Kozuka-Hata, H., Oyama, M., et al. (2012). F1Fo-ATPase, F-type proton-translocating ATPase, at the plasma membrane is critical for efficient influenza virus budding. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(12), 4615–20. https://doi.org/10.1073/pnas.1114728109

Graef, K. M., Vreede, F. T., Lau, Y.-F., McCall, A. W., Carr, S. M., Subbarao, K., et al. (2010). The PB2 subunit of the influenza virus RNA polymerase affects virulence by interacting with the mitochondrial antiviral signaling protein and inhibiting expression of beta interferon. *Journal of Virology*, *84*(17), 8433–45. https://doi.org/10.1128/JVI.00879-10

Guex, N., & Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, *18*(15), 2714–23. https://doi.org/10.1002/elps.1150181505

Guilligay, D., Tarendeau, F., Resa-Infante, P., Coloma, R., Crepin, T., Sehr, P., et al. (2008). The structural basis for cap binding by influenza virus polymerase subunit PB2. *Nature Structural & Molecular Biology*, *15*(5), 500–506. https://doi.org/10.1038/nsmb.1421

Guvench, O., & MacKerell, A. D. (2008). Comparison of Protein Force Fields for Molecular Dynamics Simulations. In A. Kukol (Ed.), *Molecular Modeling of Proteins* (1st ed., Vol. 443, pp. 63–88). Humana Press. https://doi.org/10.1007/978-1-59745-177-2_4

Hale, B. G. (2014). Conformational plasticity of the influenza A virus NS1 protein. *The Journal of General Virology*, 2099–2105. https://doi.org/10.1099/vir.0.066282-0

Hale, B. G., Randall, R. E., Ortin, J., & Jackson, D. (2008). The multifunctional NS1 protein of influenza A viruses. *Journal of General Virology*, *89*, 2359–2376. https://doi.org/10.1099/vir.0.2008/004606-0

Hale, B. G., Steel, J., Manicassamy, B., Medina, R. A., Ye, J., Hickman, D., et al. (2010). Mutations in the NS1 C-terminal tail do not enhance replication or virulence of the 2009 pandemic H1N1 influenza A virus. *The Journal of General Virology*, *91*(Pt 7), 1737–42. https://doi.org/10.1099/vir.0.020925-0

Hall, D. R., Kozakov, D., & Vajda, S. (2012). Analysis of protein binding sites by computational solvent mapping. In R. Baron (Ed.), *Computational Drug Discovery and Design. Methods in Molecular Biology (Methods and Protocols).* (Vol. 819, pp. 13–27). Springer New York, NY. https://doi.org/10.1007/978-1-61779-465-0_2

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*. https://doi.org/citeulike-article-id:691774

Hatakeyama, D., Shoji, M., Yamayoshi, S., Hirota, T., Nagae, M., Yanagisawa, S., et al. (2014). A novel functional site in the PB2 subunit of influenza A virus essential for acetyl-CoA interaction, RNA polymerase activity, and viral replication. *The Journal of Biological Chemistry*, *289*(36), 24980–94. https://doi.org/10.1074/jbc.M114.559708

Hayden, F. G., Cote, K. M., & Douglas, R. G. (1980). Plaque inhibition assay for drug susceptibility testing of influenza viruses. *Antimicrobial Agents and Chemotherapy*, *17*(5), 865–70.

Heiny, A. T., Miotto, O., Srinivasan, K. N., Khan, A. M., Zhang, G. L., Brusic, V., et al. (2007). Evolutionarily conserved protein sequences of influenza a viruses, avian and human, as vaccine targets. *PloS One*, *2*(11), e1190. https://doi.org/10.1371/journal.pone.0001190

Hess, B., Kutzner, C., van der Spoel, D., & Lindahl, E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation*, *4*(3), 435–447. https://doi.org/10.1021/ct700301q

Houston, D. R., & Walkinshaw, M. D. (2013). Consensus docking: Improving the reliability of docking in a virtual screening context. *Journal of Chemical Information and Modeling*, *53*(2), 384–390. https://doi.org/10.1021/ci300399w

Huang, N., Shoichet, B. K., & Irwin, J. J. (2006). Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry*, *49*(23), 6789–6801. https://doi.org/10.1021/jm0608356

Huang, S., Chen, J., Chen, Q., Wang, H., Yao, Y., Chen, J., et al. (2013). A second CRM1-dependent nuclear export signal in the influenza A virus NS2 protein contributes to the nuclear export of viral ribonucleoproteins. *Journal of Virology*, *87*(2), 767–78. https://doi.org/10.1128/JVI.06519-11

Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics (Oxford, England)*, *26*(5), 680–2. https://doi.org/10.1093/bioinformatics/btq003

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: visual molecular dynamics. *Journal of Molecular Graphics*, *14*(1), 33–8, 27–8.

Hurt, A. C., Holien, J. K., Parker, M. W., & Barr, I. G. (2009). Oseltamivir resistance and the H274Y neuraminidase mutation in seasonal, pandemic and highly pathogenic influenza viruses. *Drugs*, *69*(18), 2523–31. https://doi.org/10.2165/11531450-000000000-00000

Hutchinson, E. C., Charles, P. D., Hester, S. S., Thomas, B., Trudgian, D., Martínez-Alonso, M., et al. (2014). Conserved and host-specific features of influenza virion architecture. *Nature Communications*, *5*(4816). https://doi.org/10.1038/ncomms5816

Iwatsuki-Horimoto, K., Horimoto, T., Fujii, Y., & Kawaoka, Y. (2004). Generation of influenza A virus NS2 (NEP) mutants with an altered nuclear export signal sequence. *Journal of Virology*, *78*(18), 10149–55. https://doi.org/10.1128/JVI.78.18.10149-10155.2004

Jones, D. T., & Cozzetto, D. (2015). DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics (Oxford, England)*, *31*(6), 857–63. https://doi.org/10.1093/bioinformatics/btu744

Kakisaka, M., Sasaki, Y., Yamada, K., Kondoh, Y., Hikono, H., Osada, H., et al. (2015). A Novel Antiviral Target Structure Involved in the RNA Binding, Dimerization, and Nuclear Export Functions of the Influenza A Virus Nucleoprotein. *PLoS Pathogens*, *11*(7), e1005062. https://doi.org/10.1371/journal.ppat.1005062

Kalia, M., & Kukol, A. (2011). Structure and dynamics of the kinase IKK-?? - A key regulator of the NF-kappa B transcription factor. *Journal of Structural Biology*, *176*(2), 133–142. https://doi.org/10.1016/j.jsb.2011.07.012

Kao, R. Y., Yang, D., Lau, L.-S., Tsui, W. H. W., Hu, L., Dai, J., et al. (2010). Identification of influenza A nucleoprotein as an antiviral target. *Nature Biotechnology*, *28*(6), 600–605. https://doi.org/10.1038/nbt.1638

Karthikeyan, M., & Vyas, R. (2014). *Practical Chemoinformatics*. Springer Verlag. https://doi.org/10.1007/978-81-322-1780-0

Kirui, J., Bucci, M. D., Poole, D. S., & Mehle, A. (2014). Conserved features of the PB2 627 domain impact influenza virus polymerase function and replication. *Journal of Virology*, *88*(11), 5977–86. https://doi.org/10.1128/JVI.00508-14

Kitchen, D., Decornez, H., Furr, J., & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery*, *3*(11), 935–949. https://doi.org/10.1038/nrd1549

Koszalka, P., Tilmanis, D., & Hurt, A. C. (2017). Influenza antivirals currently in late-phase clinical trial. *Influenza and Other Respiratory Viruses*. https://doi.org/10.1111/irv.12446

Krug, R. M. (2015). Functions of the influenza A virus NS1 protein in antiviral defense. *Current Opinion in Virology*, *12*, 1–6. https://doi.org/10.1016/j.coviro.2015.01.007

Kukol, A. (2011). Consensus virtual screening approaches to predict protein ligands. *European Journal of Medicinal Chemistry*, *46*(9), 4661–4664. https://doi.org/10.1016/j.ejmech.2011.05.026

Kukol, A., & Hughes, D. J. (2014). Large-scale analysis of influenza A virus nucleoprotein sequence conservation reveals potential drug-target sites. *Virology*, *454–455*(1), 40–47.

Kukol, A., & Patel, H. (2014). Influenza A nucleoprotein binding sites for antivirals : current research and future potential. *Future Virology*, *9*, 625–627.

Kuzuhara, T., Kise, D., Yoshida, H., Horika, T., Murazaki, Y., Nishimura, A., et al. (2009). Structural basis of the influenza A virus RNA polymerase PB2 RNA-binding domain containing the pathogenicity-determinant lysine 627 residue. *Journal of Biological Chemistry*, *284*(11), 6855–6860. https://doi.org/10.1074/jbc.C800224200

Lagorce, D., Sperandio, O., Baell, J. B., Miteva, M. A., & Villoutreix, B. O. (2015). FAF-Drugs3: A web server for compound property calculation and chemical library design. *Nucleic Acids Research*, *43*(W1). https://doi.org/10.1093/nar/gkv353

Laskowski, R. A., & Swindells, M. B. (2011). LigPlot+: Multiple Ligand–Protein Interaction Diagrams for Drug Discovery. *Journal of Chemical Information and Modeling*, *51*(10), 2778–2786. https://doi.org/10.1021/ci200227u

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., et al. (2014). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research*, *42*(Database issue), D1091-7. https://doi.org/10.1093/nar/gkt1068

Le Gall, T., Romero, P. R., Cortese, M. S., Uversky, V. N., & Dunker, A. K. (2007). Intrinsic disorder in the Protein Data Bank. *Journal of Biomolecular Structure & Dynamics*, *24*(4), 325–42. https://doi.org/10.1080/07391102.2007.10507123

Lejal, N., Tarus, B., Bouguyon, E., Chenavas, S., Bertho, N., Delmas, B., et al. (2013). Structure-based discovery of the novel antiviral properties of naproxen against the nucleoprotein of influenza A virus. *Antimicrobial Agents and Chemotherapy*, *57*(5), 2231–42. https://doi.org/10.1128/AAC.02335-12

Li, W., Noah, J. W., & Noah, D. L. (2011). Alanine substitutions within a linker region of the influenza A virus non-structural protein 1 alter its subcellular localization and attenuate virus replication. *Journal of General Virology*, *92*(8), 1832–1842. https://doi.org/10.1099/vir.0.031336-0

Li, X., Li, Y., Cheng, T., Liu, Z., & Wang, R. (2010). Evaluation of the performance of four molecular docking programs on a diverse set of protein-ligand complexes. *Journal of Computational Chemistry*, *31*(11), 2109–2125. https://doi.org/10.1002/jcc.21498

Lin, D., Lan, J., & Zhang, Z. (2007). Structure and function of the NS1 protein of influenza A virus. *Acta Biochimica et Biophysica Sinica*, *39*(3), 135–162. https://doi.org/10.1111/j.1745-7270.2007.00263.x

Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, *46*(1), 3–26. https://doi.org/10.1016/S0169-409X(00)00129-0

Liu, X., Ouyang, S., Yu, B., Liu, Y., Huang, K., Gong, J., et al. (2010). PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. *Nucleic Acids Research*, *38*(Web Server issue), W609-14. https://doi.org/10.1093/nar/gkq300

Liu, Y., Qin, K., Meng, G., Zhang, J., Zhou, J., Zhao, G., et al. (2013). Structural and functional characterization of K339T substitution identified in the PB2 subunit cap-binding pocket of influenza A virus. *Journal of Biological Chemistry*, *288*(16), 11013–11023. https://doi.org/10.1074/jbc.M112.392878

Lomenick, B., Hao, R., Jonai, N., Chin, R. M., Aghajan, M., Warburton, S., et al. (2009). Target identification using drug affinity responsive target stability (DARTS). *Proceedings of the National Academy of Sciences of the United States of America*, *106*(51), 21984–9. https://doi.org/10.1073/pnas.0910040106

Lomenick, B., Jung, G., Wohlschlegel, J. A., & Huang, J. (2011). Target identification using drug affinity responsive target stability (DARTS). *Current Protocols in Chemical Biology*, *4*(3), 163–180. https://doi.org/10.1002/9780470559277.ch110180

Lommer, B. S., & Luo, M. (2002). Structural plasticity in influenza virus protein NS2 (NEP). *The Journal of Biological Chemistry*, *277*(9), 7108–17. https://doi.org/10.1074/jbc.M109045200

Long, J. S., Giotis, E. S., Moncorgé, O., Frise, R., Mistry, B., James, J., et al. (2016). Species difference in ANP32A underlies influenza A virus polymerase host restriction. *Nature*, *529*(7584), 101–104. https://doi.org/10.1038/nature16474

Mänz, B., Brunotte, L., Reuther, P., & Schwemmle, M. (2012). Adaptive mutations in NEP compensate for defective H5N1 RNA replication in cultured human cells. *Nature Communications*, *3*(May), 802. https://doi.org/10.1038/ncomms1804

Mehle, A., & Doudna, J. A. (2009). Adaptive strategies of the influenza virus polymerase for replication in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(50), 21312–6. https://doi.org/10.1073/pnas.0911915106

Mibayashi, M., Martínez-Sobrido, L., Loo, Y.-M., Cárdenas, W. B., Gale, M., García-Sastre, A., et al. (2007). Inhibition of retinoic acid-inducible gene I-mediated induction of beta interferon by the NS1 protein of influenza A virus. *Journal of Virology*, *81*(2), 514–24. https://doi.org/10.1128/JVI.01265-06

Min, J.-Y., Li, S., Sen, G. C., & Krug, R. M. (2007). A site on the influenza A virus NS1 protein mediates both inhibition of PKR activation and temporal regulation of viral RNA synthesis. *Virology*, *363*(1), 236–43.

https://doi.org/10.1016/j.virol.2007.01.038

Mitnaul, L. J., Matrosovich, M. N., Castrucci, M. R., Tuzikov, A. B., Bovin, N. V, Kobasa, D., et al. (2000). Balanced hemagglutinin and neuraminidase activities are critical for efficient replication of influenza A virus. *Journal of Virology*, *74*(13), 6015–20.

Moncorgé, O., Mura, M., & Barclay, W. S. (2010). Evidence for avian and human host cell factors that affect the activity of influenza virus polymerase. *Journal of Virology*, *84*(19), 9978–86. https://doi.org/10.1128/JVI.01134-10

Moriyama, M., Chen, I.-Y., Kawaguchi, A., Koshiba, T., Nagata, K., Takeyama, H., et al. (2016). The RNA- and TRIM25-Binding Domains of Influenza Virus NS1 Protein Are Essential for Suppression of NLRP3 Inflammasome-Mediated Interleukin-1β Secretion. *Journal of Virology*, *90*(8), 4105–14. https://doi.org/10.1128/JVI.00120-16

Morris, G., & Huey, R. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, *30*(16), 2785–2791. https://doi.org/10.1002/jcc.21256.AutoDock4

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., et al. (1998). Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function, *19*(14), 1639–1662.

Moscona, A. (2009). Global Transmission of Oseltamivir-Resistant Influenza. *New England Journal of Medicine*, *360*(10), 953–956. https://doi.org/10.1056/NEJMp0900648

Naesens, L., Stevaert, A., & Vanderlinden, E. (2016). Antiviral therapies on the horizon for influenza. *Current Opinion in Pharmacology*, *30*, 106–115. https://doi.org/10.1016/j.coph.2016.08.003

Neumann, G., Hughes, M. T., & Kawaoka, Y. (2000). Influenza A virus NS2 protein mediates vRNP nuclear export through NES-independent interaction with hCRM1. *The EMBO Journal*, *19*(24), 6751–8. https://doi.org/10.1093/emboj/19.24.6751

Neumann, G., & Kawaoka, Y. (2015). Transmission of influenza A viruses. *Virology*, *479*, 234–246. https://doi.org/10.1016/j.virol.2015.03.009

Ng, A. K.-L., Chan, W.-H., Choi, S.-T., Lam, M. K.-H., Lau, K.-F., Chan, P. K.-S., et al. (2012). Influenza polymerase activity correlates with the strength of interaction between nucleoprotein and PB2 through the host-specific residue K/E627. *PloS One*, *7*(5), e36415. https://doi.org/10.1371/journal.pone.0036415

Nicholls, A. (2008). What do we know and when do we know it? *Journal of Computer-Aided Molecular Design*, *22*(3–4), 239–255. https://doi.org/10.1007/s10822-008-9170-2

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, *3*(1), 33. https://doi.org/10.1186/1758-2946-3-33

O'Neill, R. E., Talon, J., & Palese, P. (1998). The influenza virus NEP (NS2 protein) mediates the nuclear export of viral ribonucleoproteins. *EMBO Journal*, *17*(1), 288–296. https://doi.org/10.1093/emboj/17.1.288

Oliveira, S. H. P., Ferraz, F. A. N., Honorato, R. V, Xavier-Neto, J., Sobreira, T. J. P., & de Oliveira, P. S. L. (2014). KVFinder: steered identification of protein cavities as a PyMOL plugin. *BMC Bioinformatics*, *15*, 197. https://doi.org/10.1186/1471-2105-15-197

Ortigoza, M. B., Dibben, O., Maamary, J., Martinez-Gil, L., Leyva-Grado, V. H., Abreu, P., et al. (2012). A novel small molecule inhibitor of influenza A viruses that targets polymerase function and indirectly induces interferon. *PLoS Pathogens*, *8*(4). https://doi.org/10.1371/journal.ppat.1002668

Patel, H., & Kukol, A. (2016a). Evaluation of a novel virtual screening strategy using receptor decoy binding sites. *Journal of Negative Results in BioMedicine*, *15*(15). https://doi.org/10.1186/s12952-016-0058-8

Patel, H., & Kukol, A. (2016b). Recent discoveries of influenza A drug target sites to combat virus replication. *Biochemical Society Transactions*, *44*(3), 932–936. https://doi.org/10.1042/BST20160002

Patel, H., & Kukol, A. (2017). Evolutionary conservation of influenza A PB2 sequences reveals potential target sites for small molecule inhibitors. *Virology*, *509*, 112–120. https://doi.org/10.1016/j.virol.2017.06.009

Paterson, D., & Fodor, E. (2012). Emerging roles for the influenza A virus nuclear export protein (NEP). *PLoS Pathogens*, *8*(12), e1003019. https://doi.org/10.1371/journal.ppat.1003019

Pautus, S., Sehr, P., Lewis, J., Fortune, A., Wolkerstorfer, A., Szolar, O., et al. (2013). New 7 ‐ Methylguanine Derivatives Targeting the Influenza Polymerase PB2 Cap-Binding Domain. *Journal of Medicinal Chemistry*, *56*, 8915–8930.

Pflug, A., Guilligay, D., Reich, S., & Cusack, S. (2014). Structure of influenza A polymerase bound to the viral RNA promoter. *Nature*, *516*(7531), 355–60. https://doi.org/10.1038/nature14008

Pflug, A., Lukarska, M., Resa-Infante, P., Reich, S., & Cusack, S. (2017). Structural insights into RNA synthesis by the influenza virus transcription-replication machine. *Virus Research*. https://doi.org/10.1016/j.virusres.2017.01.013

Plewczynski, D., Łaźniewski, M., Augustyniak, R., & Ginalski, K. (2011). Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *Journal of Computational Chemistry*, *32*(4), 742–55. https://doi.org/10.1002/jcc.21643

Poole, E., Elton, D., Medcalf, L., & Digard, P. (2004). Functional domains of the influenza A virus PB2 protein: identification of NP- and PB1-binding sites. *Virology*, *321*(1), 120–33. https://doi.org/10.1016/j.virol.2003.12.022

Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., et al. (2013). GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, *29*(7), 845–854. https://doi.org/10.1093/bioinformatics/btt055

Reich, S., Guilligay, D., Pflug, A., Malet, H., Berger, I., Crépin, T., et al. (2014). Structural insight into cap-snatching and RNA synthesis by influenza polymerase. *Nature*, *516*(7531), 361–6. https://doi.org/10.1038/nature14009

Reuther, P., Giese, S., Götz, V., Kilb, N., Mänz, B., Brunotte, L., et al. (2014). Adaptive mutations in the nuclear export protein of human-derived H5N1 strains facilitate a polymerase activity-enhancing conformation. *Journal of Virology*, *88*(1), 263–71. https://doi.org/10.1128/JVI.01495-13

Reuther, P., Giese, S., Gotz, V., Riegger, D., & Schwemmle, M. (2014). Phosphorylation of Highly Conserved Serine Residues in the Influenza A Virus Nuclear Export Protein NEP Plays a Minor Role in Viral Growth in Human Cells and Mice. *Journal of Virology*, *88*(13), 7668–7673.

Richardson, J. C., & Akkina, R. K. (1991). NS2 protein of influenza virus is found in purified virus and phosphorylated in infected cells. *Archives of Virology*, *116*(1–4), 69–80. https://doi.org/10.1007/BF01319232

Robb, N. C., Smith, M., Vreede, F. T., & Fodor, E. (2009). NS2/NEP protein regulates transcription and replication of the influenza virus RNA genome. *The Journal of General Virology*, *90*(Pt 6), 1398–407. https://doi.org/10.1099/vir.0.009639-0

Samson, M., Pizzorno, A., Abed, Y., & Boivin, G. (2013). Influenza virus resistance to neuraminidase inhibitors. *Antiviral Research*, *98*(2), 174–85. https://doi.org/10.1016/j.antiviral.2013.03.014

Sayle, R. A., & Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all. *Trends in Biochemical Sciences*, *20*(9), 374.

Schnell, J. R., & Chou, J. J. (2008). Structure and mechanism of the M2 proton channel of influenza A virus. *Nature*, *451*(7178), 591–5. https://doi.org/10.1038/nature06531

Shi, Y., Wu, Y., Zhang, W., Qi, J., & Gao, G. F. (2014). Enabling the “host jump”: structural determinants of receptor-binding specificity in influenza A viruses. *Nature Reviews. Microbiology*, *12*(12), 822–31. https://doi.org/10.1038/nrmicro3362

Shimizu, T., Takizawa, N., Watanabe, K., Nagata, K., & Kobayashi, N. (2011). Crucial role of the influenza virus NS2 (NEP) C-terminal domain in M1 binding and nuclear export of vRNP. *FEBS Letters*, *585*(1), 41–6. https://doi.org/10.1016/j.febslet.2010.11.017

Sidwell, R. W., & Smee, D. F. (2000). In vitro and in vivo assay systems for study of influenza virus inhibitors. *Antiviral Research*, *48*(1), 1–16. https://doi.org/10.1016/S0166-3542(00)00125-X

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*(539). https://doi.org/10.1038/msb.2011.75

Stevaert, A., & Naesens, L. (2016). The Influenza Virus Polymerase Complex: An Update on Its Structure, Functions, and Significance for Antiviral Drug Design. *Medicinal Research Reviews*, 1–47. https://doi.org/10.1002/med.21401

Stiver, G. (2003). The treatment of influenza with antiviral drugs. *CMAJ : Canadian Medical Association Journal = Journal de l’Association Medicale Canadienne*, *168*(1), 49–56.

Sugiyama, K., Obayashi, E., Kawaguchi, A., Suzuki, Y., Tame, J. R. H., Nagata, K., et al. (2009). Structural insight into the essential PB1-PB2 subunit contact of the influenza virus RNA polymerase. *The EMBO Journal*, *28*(12), 1803–11. https://doi.org/10.1038/emboj.2009.138

Talon, J., Horvath, C. M., Polley, R., Basler, C. F., Muster, T., Palese, P., et al. (2000). Activation of interferon regulatory factor 3 is inhibited by the influenza A virus NS1 protein. *Journal of Virology*, *74*(17), 7989–96.

Taubenberger, J. K., & Kash, J. C. (2010). Influenza virus evolution, host adaptation, and pandemic formation. *Cell Host and Microbe*, *7*(6), 440–451. https://doi.org/10.1016/j.chom.2010.05.009

Taubenberger, J. K., & Morens, D. M. (2008). The pathology of influenza virus infections. *Annual Review of Pathology*, *3*, 499–522. https://doi.org/10.1146/annurev.pathmechdis.3.121806.154316

Thierry, E., Guilligay, D., Kosinski, J., Bock, T., Gaudon, S., Round, A., et al. (2016). Influenza Polymerase Can Adopt an Alternative Configuration Involving a Radical Repacking of PB2 Domains. *Molecular Cell*, *61*(1), 125–137. https://doi.org/10.1016/j.molcel.2015.11.016

Tong, S., Zhu, X., Li, Y., Shi, M., Zhang, J., Bourgeois, M., et al. (2013). New world bats harbor diverse influenza A viruses. *PLoS Pathogens*, *9*(10), e1003657. https://doi.org/10.1371/journal.ppat.1003657

Trott, O., & Olson, A. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *Journal of Computational Chemistry*, *31*(2), 455–461. https://doi.org/10.1002/jcc.21334.AutoDock

Tu, J., Guo, J., Zhang, A., Zhang, W., Zhao, Z., Zhou, H., et al. (2011). Effects of the C-Terminal Truncation in NS1 Protein of the 2009 Pandemic H1N1 Influenza Virus on Host Gene Expression. *PLoS ONE*, *6*(10), e26175. https://doi.org/10.1371/journal.pone.0026175

Tuccinardi, T., Poli, G., Romboli, V., Giordano, A., & Martinelli, A. (2014). Extensive consensus docking evaluation for ligand pose prediction and virtual screening studies. *Journal of Chemical Information and Modeling*, *54*(10), 2980–6. https://doi.org/10.1021/ci500424n

Twu, K. Y., Noah, D. L., Rao, P., Kuo, R.-L., & Krug, R. M. (2006). The CPSF30 binding site on the NS1A protein of influenza A virus is a potential antiviral target. *Journal of Virology*, *80*(8), 3957–65. https://doi.org/10.1128/JVI.80.8.3957-3965.2006

Tynell, J., Melén, K., & Julkunen, I. (2014). Mutations within the conserved NS1 nuclear export signal lead to inhibition of influenza A virus replication. *Virology Journal*, *11*(1), 128. https://doi.org/10.1186/1743-422X-11-128

Valdar, W. S. J. (2002). Scoring residue conservation. *Proteins: Structure, Function and Genetics*, *48*(October 2001), 227–241. https://doi.org/10.1002/prot.10146

Vasin, A. V, Temkina, O. A., Egorov, V. V, Klotchenko, S. A., Plotnikova, M. A., & Kiselev, O. I. (2014). Molecular mechanisms enhancing the proteome of influenza A viruses: An overview of recently discovered proteins. *Virus Research*, *185*, 53–63. https://doi.org/10.1016/j.virusres.2014.03.015

Wang, J., Ma, C., Fiorin, G., Carnevale, V., Wang, T., Hu, F., et al. (2011). Molecular dynamics simulation directed rational design of inhibitors targeting drug-resistant mutants of influenza A virus M2. *Journal of the American Chemical Society*, *133*(32), 12834–41. https://doi.org/10.1021/ja204969m

Wang, R., Lu, Y., Fang, X., & Wang, S. (2004). An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. *Journal of Chemical Information and Computer Sciences*, *44*(6), 2114–2125. https://doi.org/10.1021/ci049733j

Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F., & Jones, D. T. (2004). The DISOPRED server for the prediction of protein disorder. *Bioinformatics (Oxford, England)*, *20*(13), 2138–9. https://doi.org/10.1093/bioinformatics/bth195

Warren, S., Wan, X. F., Conant, G., & Korkin, D. (2013). Extreme evolutionary conservation of functionally important regions in H1N1 influenza proteome. *PLoS ONE*, *8*(11), 1–14. https://doi.org/10.1371/journal.pone.0081027

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics (Oxford, England)*, *25*(9), 1189–91. https://doi.org/10.1093/bioinformatics/btp033

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., et al. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, *34*(90001), D668–D672. https://doi.org/10.1093/nar/gkj067

Wu, N. C., Olson, C. A., Du, Y., Le, S., Tran, K., Remenyi, R., et al. (2015).

Functional Constraint Profiling of a Viral Protein Reveals Discordance of Evolutionary Conservation and Functionality. *PLoS Genetics*, *11*(7), 1–27. https://doi.org/10.1371/journal.pgen.1005310

Wu, Y., Wu, Y., Tefsen, B., Shi, Y., Gao, G. F., Palese, P., et al. (2014). Bat-derived influenza-like viruses H17N10 and H18N11. *Trends in Microbiology*, *22*(4), 183–91. https://doi.org/10.1016/j.tim.2014.01.010

Xu, C., Hu, W.-B., Xu, K., He, Y.-X., Wang, T.-Y., Chen, Z., et al. (2012). Amino acids 473V and 598P of PB1 from an avian-origin influenza A virus contribute to polymerase activity, especially in mammalian cells. *The Journal of General Virology*, *93*(Pt 3), 531–40. https://doi.org/10.1099/vir.0.036434-0

Yamada, S., Hatta, M., Staker, B. L., Watanabe, S., Imai, M., Shinya, K., et al. (2010). Biological and structural characterization of a host-adapting amino acid in influenza virus. *PLoS Pathogens*, *6*(8), e1001034. https://doi.org/10.1371/journal.ppat.1001034

Yang, J., & Zhang, Y. (2015). Protein Structure and Function Prediction Using I-TASSER. *Current Protocols in Bioinformatics*, *52*, 5.8.1-15. https://doi.org/10.1002/0471250953.bi0508s52

Yuan, S., Chu, H., Ye, J., Singh, K., Ye, Z., Zhao, H., et al. (2017). Identification of a novel small-molecule compound targeting the influenza A virus polymerase PB1-PB2 interface. *Antiviral Research*, *137*, 58–66. https://doi.org/10.1016/j.antiviral.2016.11.005

Zambon, M. (1998). Laboraotry Diagnosis of Influenza. In K. G. Nicholson, R. Webster, & A. J. Hay (Eds.), *Textbook of Influenza* (First, pp. 291–313). Blackwell Science.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, *9*, 40. https://doi.org/10.1186/1471-2105-9-40

Zhu, Z., Shi, Z., Yan, W., Wei, J., Shao, D., Deng, X., et al. (2013). Nonstructural Protein 1 of Influenza A Virus Interacts with Human Guanylate-Binding Protein 1 to Antagonize Antiviral Activity. *PLoS ONE*, *8*(2), e55920. https://doi.org/10.1371/journal.pone.0055920

9 APPENDIX

9.1 NS1 amino acid conservation scores

Table 18. Conservation scores for each amino acid from a multiple sequence alignment of NS1 protein sequences. The scores were obtained from the Jalview AAcons web server using Valdar's scoring method.

Conservation Score (Valdar)	Residue position
0.900 - 1.00 (total 155)	Met1-Val6, ser8, Phe9-Cys13, Phe14, Leu15-His17, Val18-lys20, Leu27, Asp29, Ala30-Phe32, Asp34-Arg44, Gly45-gly47, Thr49-Asp53, Ile54, Ala57, Thr58, Gly61, Lys62, Ile64-glu66, Ile68, leu69, glu72, lys75, Pro80, arg83, Tyr84, Asp87-Thr89, Glu91-Trp97, Met99-Lys105, Gly108, leu110, Ile112-Lys121, Ile123, Leu125-Val131, Phe133, Arg135-Glu137, Leu139, Ile140, Leu142-Ser160, Pro162-Thr165, Glu167-Ala172, Ile173-Asn183, Asn185,-Arg188, Thr190, Glu191, Gln194-Trp198, Glu203, Gly205, Pro211
0.800-0.899 (total 36)	Ser7, Arg21-Asp24, Gly28, Leu33, Thr56, Gln63, Gly71, Ser73, Asp74, Leu77, Lys78, Ala81, Ser82, Leu85, Thr86, Leu90, Phe98, Val106, Cys111, Ile132, Asp134, Thr138, Leu141, Leu161, Asp184, Val189, Ile193, Arg199, Asp202, Arg206-Leu209
0.700 - 0.799 (total 18)	Gln25, Glu26, Asn48, Glu55, Glu70, Ala76, Met79, Ser109, Thr122, Ile124, Thr192, Asn200, Asp204, Pro210, Asn212-Lys214
0.600-0.699	Ala60, Arg67, Ala107, Gly166, Ser201
<0.60	Arg59, Arg215

The H5N1 A/Vietnam/1203/2004 sequence numbering convention has been used where the 5 amino acid deletion is not included.

9.2 NEP amino acid conservation scores

Table 19. Conservation score for each amino acid from a multiple sequence alignment of NEP sequences. The scores were obtained from the Jalview AAcons web server using Valdar's scoring method.

Valdar conservation score	**Residue position**
0.90 - 1.00	Met1, Asp2, Pro3, Asn4, Thr5, Ser8, Phe9, Gln10, Asp11, Ile12, Leu13, Arg15, Met16, Ser17, Lys18, Met19, Gln20, Leu21, Ser23, Ser24, Ser25, Asp27, Leu28, Asn29, Gly30, Met31, Ile32, Thr33, Phe35, Glu36, Leu38, Lys39, Tyr41, Arg42, Asp43, Ser44, Leu45, Gly46, Glu47, Met50, Arg51, Gly53, Asp54, Leu55, His56, Leu58, Gln59, Arg61, Asn62, Trp65, Arg66, Glu67, Gln68, Leu69, Gln71, Lys72, Phe73, Glu74, Glu75, Ile76, Arg77, Trp78, Leu79, Ile80, Glu82, Arg84, Leu87, Thr90, Glu91, Asn92, Ser93, Phe94, Glu95, Gln96, Ile97, Thr98, Phe99, Met100, Gln101, Ala102, Leu103, His104, Leu105, Leu106, Leu107, Glu108, Val109, Glu110, Glu112, Ile113, Arg114, Phe116, Ser117, Phe118, Gln119, Leu120, Ile121
0.80-0.899	Val6, Ser7, Gln34, Ser37, Leu40, Val49, Met52, Lys64, Glu81, Val83, His85, Lys86, Lys88, Gln111, Thr115,
0.70 - 0.799	Glu22, Gly26, Ala48, Ser57, Gly70
0.60-0.699	Leu14, Asn60, Glu63, Ile89

9.3 PB2 amino acid conservation scores

Table 20. Conservation scores for each amino acid from a multiple sequence alignment of PB2 protein sequences. The scores were obtained from the Jalview AAcons web server using Valdar's scoring method.

Valdar conservation score	Residue position
0.950-1.00	Met1-Arg8, Leu10-Pro43, Leu45-Asp60, Arg62, Ile63, Ile67, Pro68, Glu69, Arg70, Asn71, Glu72, Gln73, Gly74, Gln75, Leu77, Trp78, Ser79, Lys80, Asp83, Ala84, Gly85, Ser86, Asp87-Arg101, Gly103, Pro104, Ala108-Val145, Asp146, Asn148, Pro149, Gly150, His151, Ala152, Asp153, Leu154, Ser155, Val172-Leu183, Ile185-Gln194, Cys196, Lys197, Ile198, Pro200, Leu201, Met202, Val203, Ala204, Tyr205-Val220, Gly222, Gly223, Thr224, Ser226-Gly247, Gly248, Val250, Asn252-Ala270, Val272-His285, Thr287-Gly291, Arg293-Leu298, Gln300-Glu305, Ala307-Thr333, Gly335, Ser336, Ser337, Glu341, Glu342, Glu343, Leu345-Ile354, Val356-Gly367, Arg369-Leu445, Gln447-Gly450, Glu452, Ile454, Asp455, Val457-Glu472, Ser474, Arg476, Gly477, Arg479-Lys482, Gly484-Val494, Val496-Gln507, Gly509-Gly523, Glu525-Glu558, Lys561-Gln566, Pro568, Thr569, Leu571, Tyr572, Asn573, Lys574, Glu576-Ala587, Arg589, Tyr592-Arg597, Leu599-Thr612, Gln614-Pro626, Gln628-Gly644, Arg646, Ile647, Leu648, Arg650-Lys660, Thr662-Thr666, Leu668-Gly673, Leu675, Glu677-Gly682, Gly685-Glu700, Lys702-Ile710, Glu712-Ser714, Leu716, Lys718-Arg755
0.900-0.949	Asp9, Ala44, Lys61, Glu65, Met66, Thr81, Asn102, Thr106, Ala156, Lys157-Glu171, Ala199, Ala221, Glu249, Arg251, Ser286, Arg299, Gln306, Ser334, Val338, Val344, Arg355, Arg368, Phe446, Ile451, Leu475, Val478, Val495, Thr524, Val560, Asp567, Met570, Met575, Thr598, Met645, Val649, Val667, Ala674, Thr683, Asp701, Asn711, Asn715, Ala717, Ala757, Asn759
0.850-0.899	Ile64, Thr76, Asn82, Ala105, Thr184, Asp195, Ser225, Lys339, Lys340, Pro453, Asn456, Met473, Arg508, Gln591, Val613, Lys627, Ala684, Met756, Ile758
0.800-0.849	Ser107, Thr271, Ile292, Met483, Ala661, Thr676
0.700-0.799	Ile147, Thr559, Ala588, Gly590

9.4 Molecular dynamics parameter file for production run (NS1)

```
title           = part free MD 100 ns

integrator      = md                ; leap-frog Algorithm
nsteps          = 50000000          ; 0.002*50000000 = 100 ns
dt              = 0.002             ; 2 femtosecond time step
nstenergy       = 1000              ; save energies every 5 ps
nstlist         = 10                ; frequency to update the neighbor list
rlist           = 1.0               ; short range neighbor list cut-off (in nm)
coulombtype     = pme               ; treatment of long range electrostatics
rcoulomb        = 1.0               ; short range electrostatic cut-off (in nm)
rvdw            = 1.0               ; short range van der Waals cut-off (in nm)
vdw-type        = cut-off

continuation    = yes               ; continue simulation from equilibration stage

nstxout         = 0                 ; frequency to save output coordinates
nstvout         = 0                 ; frequency to save output velocities
nstfout         = 0                 ; frequency to save output forces
nstlog          = 10000             ; update log file every 5 ps
nstxtcout       = 10000             ; compressed trajectory output every 20 ps


tcoupl          = v-rescale         ; modified Berendson thermostat
tc-grps         = protein water_and_ions     ; groups to couple separately
tau-t           = 0.5 0.5           ; time constant for coupling (in ps)
ref-t           = 300 300           ; reference temperature for coupling (K)
pcoupl          = parrinello-rahman     ; pressure coupling is on
pcoupltype      = isotropic         ; uniform scaling of box vectors
tau-p           = 2.0               ; time constant for coupling (in ps)
compressibility = 4.5e-5            ; isothermal compressibility of water (bar$^{-1}$)
ref-p           = 1.0               ; reference pressure for coupling (bar)


define          = -DPARTIAL         ; include partial position restraints

refcoord_scaling = com             ; scale centre of mass of reference coordinates
constraints     = all-bonds         ; all bonds treated as constraints

gen_vel          = yes              ; assign velocities from Maxwell distribution
```

9.5 PERL script to extract compounds within a 80% similarity cut-off

```
#!/usr/bin/perl

############################################################
# Usage: extract.pl <file with ZINC codes> <sdf-file>
#
#  prints out the mol2-entry corresponding to each ZINC code
#  contained in the file with ZINC codes
#
# (c) A.K. 20/01/2015, University of Hertfordshire
############################################################

if ($#ARGV<1) {
   print "----------------------------------------------------------------
\n";
   print "Usage: extract.pl <file with ZINC codes> <mol2-file>\n";
   print "Prints out the mol2-entry corresponding to each ZINC code\n";
   print "To redirect the output into a file use: > file.mol2\n";
   print "(c) A.K. 20/01/2015, University of Hertfordshire\n";
   print "----------------------------------------------------------------
\n";
   exit;
}
$ZincCodeFile = $ARGV[0];
$mol2File = $ARGV[1];

# read in ZINC codes

%code = (); # initialises a 'hash' variable
open (IN, $ZincCodeFile);
while(defined($in=<IN>)) {
     if ($in =~ m/(ZINC\w+)$/) {
         $code{$1} = 1;
     }
}
close IN;

print "\@<TRIPOS>MOLECULE\n";

open (IN, $mol2File);
while(defined($in=<IN>)) {
  if ($in =~ m/(ZINC\w+)$/) {
      $zinc = $1;
      if ($code{$zinc}==1) {
          print "$zinc\n";
          $line = <IN>;
          print $line;
          until($line =~ m/TRIPOS>MOLECULE/) {
               last if !defined($line=<IN>);   # read next line and exit
loop if end of file is reached
               print "$line";   # print out lines until @<TRIPOS>MOLECULE
is reached (incl.)
          }
       }
  }
}
close IN;
```

This script was written by Dr Andreas Kukol.

9.6 AutoDock 4 grid parameter file (NS1 protein)

```
npts 50 70 80                                  # num.grid points in xyz
gridfld rec.maps.fld                           # grid_data_file
spacing 0.375                                  # spacing(A)
receptor_types A C HD N OA SA                  # receptor atom types
ligand_types A NA SA C N HD OA S Br F Cl P I  # ligand atom types
receptor rec.pdbqt                             # macromolecule
gridcenter -5.013 -16.806 -26.516              # xyz-coordinates or auto
smooth 0.5                                     # store minimum energy w/in
rad(A)
map rec.A.map                                  # atom-specific affinity map
map rec.NA.map                                 # atom-specific affinity map
map rec.SA.map                                 # atom-specific affinity map
map rec.C.map                                  # atom-specific affinity map
map rec.N.map                                  # atom-specific affinity map
map rec.HD.map                                 # atom-specific affinity map
map rec.OA.map                                 # atom-specific affinity map
map rec.S.map                                  # atom-specific affinity map
map rec.Br.map                                 # atom-specific affinity map
map rec.F.map                                  # atom-specific affinity map
map rec.Cl.map                                 # atom-specific affinity map
map rec.P.map                                  # atom-specific affinity map
map rec.I.map                                  # atom-specific affinity map
elecmap rec.e.map                              # electrostatic potential map
dsolvmap rec.d.map                             # desolvation potential map
dielectric -0.1465                             # <0, AD4 distance-
dep.diel;>0,
```

9.7 AutoDock 4 docking parameter file (NS1 protein)

```
AutoDock_parameter_version 4.2 # used by AutoDock to validate parameter set
outlev 0                       # diagnostic output level
intelec                        # calculate internal electrostatics
seed pid time                  # seeds for random generator
ligand_types A C HD N NA OA SA # atoms types in ligand
fld rec.maps.fld               # grid_data_file
map rec.A.map                  # atom-specific affinity map
map rec.C.map                  # atom-specific affinity map
map rec.HD.map                 # atom-specific affinity map
map rec.N.map                  # atom-specific affinity map
map rec.NA.map                 # atom-specific affinity map
map rec.OA.map                 # atom-specific affinity map
map rec.SA.map                 # atom-specific affinity map
elecmap rec.e.map              # electrostatics map
desolvmap rec.d.map            # desolvation map
move ZINC04627258.pdbqt        # small molecule
about 4.1391 -5.2778 -1.8688   # small molecule center
tran0 random                   # initial coordinates/A or random
axisangle0 random              # initial orientation
dihe0 random                   # initial dihedrals (relative) or random
tstep 2.0                      # translation step/A
qstep 50.0                     # quaternion step/deg
dstep 50.0                     # torsion step/deg
torsdof 4                      # torsional degrees of freedom
rmstol 2.0                     # cluster_tolerance/A
extnrg 1000.0                  # external grid energy
e0max 0.0 10000                # max initial energy; max number of retries
ga_pop_size 150                # number of individuals in population
ga_num_evals 350000            # maximum number of energy evaluations
ga_num_generations 27000       # maximum number of generations
ga_elitism 1                   # number of top individuals to survive to
next generation
ga_mutation_rate 0.02          # rate of gene mutation
ga_crossover_rate 0.8          # rate of crossover
ga_window_size 10              #
ga_cauchy_alpha 0.0            # Alpha parameter of Cauchy distribution
ga_cauchy_beta 1.0             # Beta parameter Cauchy distribution
set_ga                         # set the above parameters for GA or LGA
sw_max_its 300                 # iterations of Solis & Wets local search
sw_max_succ 4                  # consecutive successes before changing rho
sw_max_fail 4                  # consecutive failures before changing rho
sw_rho 1.0                     # size of local search space to sample
sw_lb_rho 0.01                 # lower bound on rho
ls_search_freq 0.15            # probability of performing local search on
individual
set_psw1                       # set the above pseudo-Solis & Wets
parameters
unbound_model bound            # state of unbound ligand
ga_run 50                      # do this many hybrid GA-LS runs
analysis                       # perform a ranked cluster analysis
```

9.8 Python script for virtual screening with AutoDock 4

```
import os
import time
from subprocess import Popen,PIPE

step=100

mypath=os.getcwd()
files=open('filelist').read().splitlines()
i=0
while i<len(files):
    c=0
    # test -- write the job scripts to a file but don't run them
    q=Popen('cat > test'+str(i),shell=True,stdin=PIPE).stdin
    # actual run
    # q=Popen('qsub -N AD4-'+str(i)+' -j oe -o /dev/null -l nodes=1:ppn=1
    -l walltime=33:00:00',shell=True,stdin=PIPE).stdin

    q.write('#!/bin/bash\n')
    while c<100 and i<len(files):
        f=files[i]
        q.write('cd '+mypath+'/'+f+'\n')
        q.write('/soft/autodock/autodock4 -p '+f+'.dpf -l '+f+'.dlg \n')
        i+=1
        c+=1
    q.close()

    time.sleep(2)
```

This script was written by Prof. Martin Hardcastle (University of Hertfordshire).

9.9 Python script for virtual screening with AutoDock Vina

```
import os
import time
from subprocess import Popen,PIPE

step=100

mypath=os.getcwd()
files=open('filelist').read().splitlines()
i=0
while i<len(files):
    c=0
    # test -- write the job scripts to a file but don't run them
    q=Popen('cat > test'+str(i),shell=True,stdin=PIPE).stdin
    # actual run
    # q=Popen('qsub -N dock-'+str(i)+' -j oe -o /dev/null -l nodes=1:ppn=8
        -l walltime=33:00:00',shell=True,stdin=PIPE).stdin

    q.write('#!/bin/bash\n')
    q.write('cd '+mypath+'\n')
    while c<100 and i<len(files):
        f=files[i]
        b='dock_'+f
        q.write('mkdir -p '+b+'\n')
        q.write('/soft/autodock_vina_1_1_1_linux_x86/bin/vina --config
        '+mypath+'/conf.txt --ligand '+mypath+'/'+f+' --out
        '+mypath+'/'+b+'/out.pdbqt --log '+mypath+'/'+b+'/log.txt\n')
        i+=1
        c+=1
    q.close()


    time.sleep(2)
```

This script was written by Prof. Martin Hardcastle (University of Hertfordshire).

9.10 PERL script to combine virtual screening rank lists

```
#!/usr/bin/perl


if ($#ARGV < 2) {
   print "-------------------------------------------------------------\n";
   print "Usage: JoinRankList.pl <number_of_lists> <list1> <list2> ...\n";
   print "Simple combination of two or more rank lists from virtual 
screening\n";
   print "Ligands with a higher rank (at the top of a list) preceed those 
with a lower rank.\n";
   print "All lists should be of same length\n";
   print "-------------------------------------------------------------\n";
   exit;
}
$numlists = $ARGV[0];
for $i (1..$numlists) {
    $list[$i] = $ARGV[$i];
}

# read in the lists
# AutoDock Raccoon
# "ZINC03871916_rec/ZINC03871916_rec",-13.51
# "ZINC03834067_rec/ZINC03834067_rec",-13.31
# ....
# Vina
# dock_ZINC01539579.pdbqt -13.3
# dock_ZINC01540632.pdbqt -13.1
#

# read in ranked lists

for $i (1..$numlists) {
    $rank = 0;
    open (IN, $list[$i]);
    while(defined($in=<IN>)) {
         $rank++;
         $in =~ m/^.*(ZINC[0-9]+)[-\._:\n +]/;
         $LigandRank[$i]{$1} = $rank;
         $LigandName[$i][$rank] = $1; # note that ligands names are in rank 
order
    }
    close IN;
}
$lengthOfList = $rank;

# print new list

%NewList = ();
for $i (1..$lengthOfList) {
        for $j (1..$numlists) {
            $name = $LigandName[$j][$i];
            if (not exists ($NewList{$name})) {
               print "$name\n";
               $NewList{$name} = $LigandRank[$j]{$name};
            }
        }
}
```

This script was written by Dr. Andreas Kukol

9.11 Publications

Patel, H., & Kukol, A. (2017). Evolutionary conservation of influenza A PB2 sequences reveals potential target sites for small molecule inhibitors. *Virology*, *509*, 112–120. https://doi.org/10.1016/j.virol.2017.06.009

Patel, H., & Kukol, A. (2016). Evaluation of a novel virtual screening strategy using receptor decoy binding sites. *Journal of Negative Results in Biomedicine*, *15*(1), 15. https://doi.org/10.1186/s12952-016-0058-8

Patel, H., & Kukol, A. (2016). Recent discoveries of influenza A drug target sites to combat virus replication. *Biochemical Society Transactions*, *44*(3), 932-36. http://dx.doi.org/10.1042/BST20160002

Kukol, A. & Patel, H. (2014). Influenza A nucleoprotein binding sites for antivirals: current research and future potential. *Future Virology*, 9(7), 625-27. http://dx.doi.org/10.2217/fvl.14.45